

Pragmatic Trials in Public Health and Medicine

David M. Murray, Ph.D.

Associate Director for Prevention

Director, Office of Disease Prevention

National Institutes of Health

ENAR Webinar

May 20, 2016



National Institutes of Health
Office of Disease Prevention

Disclosure

- David M. Murray, Ph.D.
- Pragmatic Trials In Public Health and Medicine
- No commercial interests
- No discussion of off-label drug usage
- No relevant financial relationships

Schedule

- Introduction 4-22
- Planning the trial 23-40
- Analysis approaches 41-62
- Power 63-75
- Alternative designs 76-89
- Examples 90-113
- Review of recent practices 114-138
- Summary 139
- References 140-149

Pragmatic vs Explanatory Trials

- First described by Schwartz & Lellouch (1967).
 - Explanatory trials test causal research hypotheses.
 - Pragmatic trials help users choose between options for care.
- Similar to efficacy and effectiveness trials (Cochrane, 1971, cited in Flay, 1986).
 - Efficacy trials evaluate an intervention under carefully controlled conditions.
 - Effectiveness trials evaluate an intervention under real-world conditions.

Pragmatic Trials

- Pragmatic trials do not require a different set of research designs, measures, analytic methods, or other procedures.
- 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.
 - Which kind of randomized trial will depend on the research question.
- Alternative designs exist and may be preferred under certain conditions.

Three Kinds of Randomized Trials

- Individually Randomized Clinical Trials (RCTs)
 - Individuals randomized to study conditions with no connection among participants after randomization
 - Most surgical and drug trials, some behavioral trials
- Individually Randomized Group Treatment Trials (IRGTs)
 - Individuals randomized to study conditions with some connection among participants after randomization
 - Many behavioral trials
- Group-Randomized Trials (GRTs)
 - Groups randomized to study conditions with some connection among participants before and after randomization
 - Many trials conducted in communities, worksites, schools, etc.

Distinguishing Characteristics

- Group-randomized trials
 - The unit of assignment is an identifiable group.
 - Different groups are allocated to each condition.
 - The units of observation are members of the groups.
 - The number of groups allocated to each condition is usually limited.
- Individually randomized group-treatment trials
 - The unit of assignment is the individual participant.
 - Participants receive some of their treatment in physical or virtual groups or through a common change agent.
 - The number of groups or change agents is usually limited.

Alternative Labels

- Group-randomized trials are also called...
 - Cluster-randomized trials.
 - They are sometimes called community trials.
 - These labels are interchangeable.
- Individually randomized clinical trials are also called....
 - Randomized clinical trials,
 - Randomized controlled trials,
 - Controlled clinical trials.
 - These labels are interchangeable.

Examples

- **Group-randomized trials: Health Care Systems Collaboratory**
 - 9 pragmatic trials conducted in collaboration with health care systems, funded as UH2/UH3 trials by a variety of NIH ICs.
 - 8 are group-randomized trials (GRT)
 - Hospital acquired infections
 - CRC screening
 - Healthcare utilization in back pain care
 - Chronic pain management
 - Mortality in dialysis patients
 - Management of PTSD in trauma patients
 - Advanced care planning in nursing homes
 - Management of multiple chronic conditions

Examples

- Group-randomized trials: Health Care Systems Collaboratory
 - Overview papers
 - Anderson ML et al., Ethical and regulatory issues of pragmatic cluster randomized trials in contemporary health systems. *Clinical Trials*. 2015;12(3):276-86. PMC4498459.
 - Johnson KE et al., A guide to research partnerships for pragmatic clinical trials. *BMJ*. 2014;349:g6826.
 - Richesson RL et al., Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *Journal of the American Medical Informatics Association*. 2013;20(e2):e226-31. PMC3861929.

Examples

- Individually randomized group treatment trials: Childhood Obesity Prevention and Treatment Research (COPTR)
 - 4 trials funded by NHLBI as U01s
 - Two prevention studies targeting young children
 - Two treatment studies targeting youth
 - All involve substantial participant interaction post-randomization
 - Overview paper
 - Pratt CA et al., Childhood Obesity Prevention and Treatment Research (COPTR): interventions addressing multiple influences in childhood and adolescent obesity. *Contemporary Clinical Trials*. 2013;36(2):406-13.

Notation

- Following Murray (1998)
 - Dependent variable (Y)
 - Condition, C_l ($l=1\dots c$), will identify the study conditions
 - Time, T_j ($j=1\dots t$), will identify the measurement occasion
 - Group, \mathbf{G}_k ($k=1\dots g$), will identify the unit of assignment
 - Member, \mathbf{M}_i ($i=1\dots m$), will identify the unit of observation
 - Covariate, X_o ($o=1\dots x$), will identify covariates
 - Random effects will be **BOLD**, fixed effects will be PLAIN

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 most IRGTs have $N < 100$.
Confounding can be more of a concern in IRGTs than in RCTs.
- Group-randomized trials
 - GRTs often involve a limited number of groups, often < 50 .
 - There may be limited opportunity for randomization to distribute potential confounders evenly.
 - Confounding is usually a concern in GRTs if G is < 50 .

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 will share some physical, geographic, social or other connection.
- The members of groups created for an IRGT will develop similar connections.
- Those connections will create a positive intraclass correlation that reflects extra variation attributable to the group.

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

- 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 (ICC) is the fraction of the total variation in the data attributable to the unit of assignment:

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

Impact on the Analysis in a GRT

- 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

- Nested factors must be 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 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.

Impact on the Analysis

- 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.

Impact on the Analysis

- 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), $ICC_x \approx ICC_y$.
- 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

- 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:

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

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.

The Need for GRTs and IRGTs

- A GRT remains the best comparative design available when the investigator wants to evaluate an intervention that...
 - operates at a group level
 - manipulates the social or physical environment
 - cannot be delivered to individuals without contamination
- An IRGT is the best comparative design when...
 - Individual randomization is possible without contamination
 - There are good reasons to deliver the intervention in groups
- The challenge is to create trials that are:
 - Rigorous enough to avoid threats to validity of the design,
 - Analyzed so as 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 analysis 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 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, including a GRT, are:
 - Randomization
 - Replication
 - Variance reduction

Potential Threats to Internal Validity

- Four primary threats:
 - 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 Limit Threats to Internal Validity

- Randomization
- A priori matching or stratification
 - Of groups in GRTs, of members in IRGTs and RCTs
- Objective measures
- Independent evaluation personnel who are blind to conditions
- Analytic strategies
 - Regression adjustment for covariates
- 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 Validity of 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.
 - Consider alternate models for time.
 - Assess potential confounding and effect modification.

Strategies to Protect the Validity of 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.
 - Employ regression adjustment for covariates to reduce variance and intraclass correlation.

Factors That Can Reduce Precision

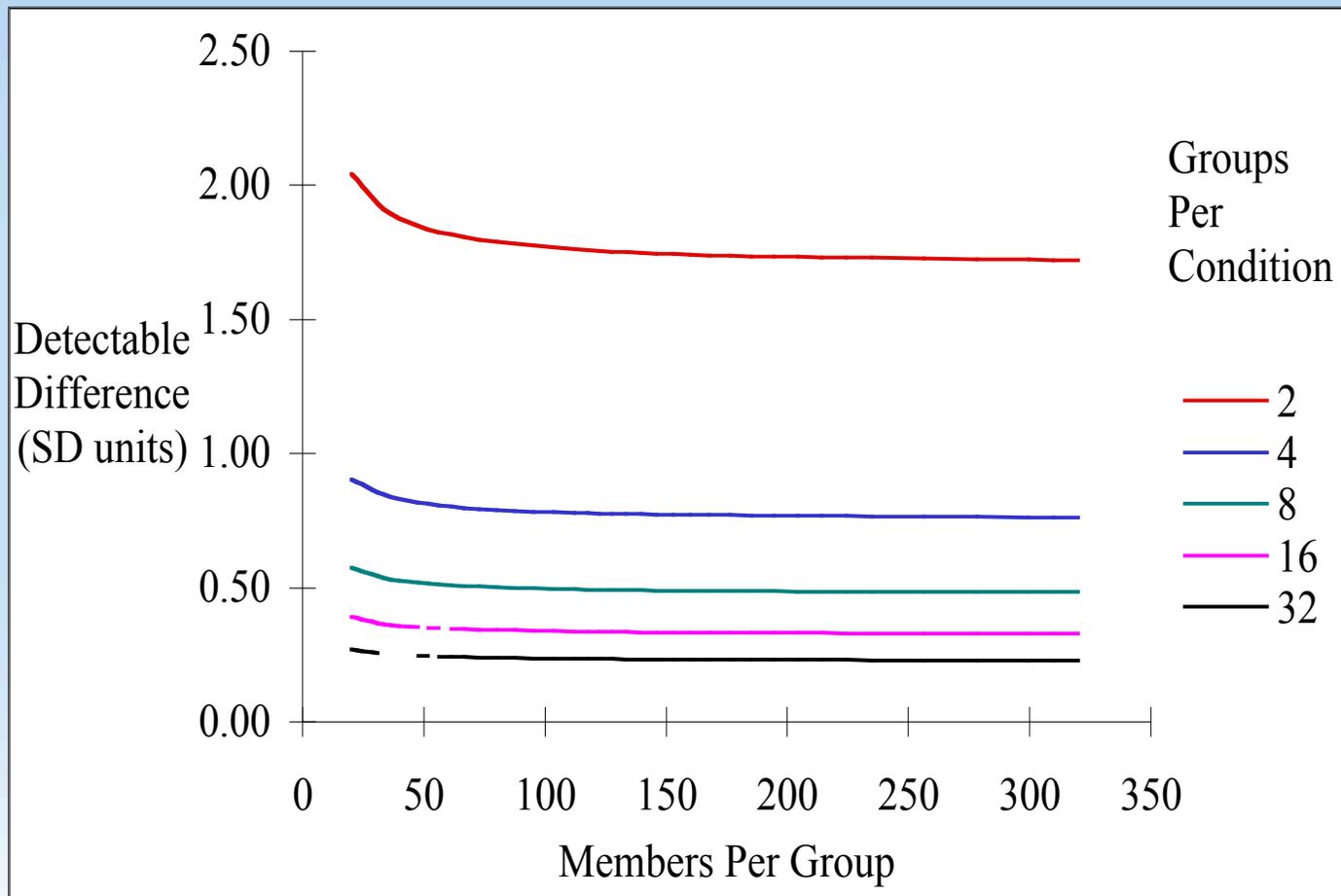
- The variance of the condition mean in a GRT is:

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

- 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

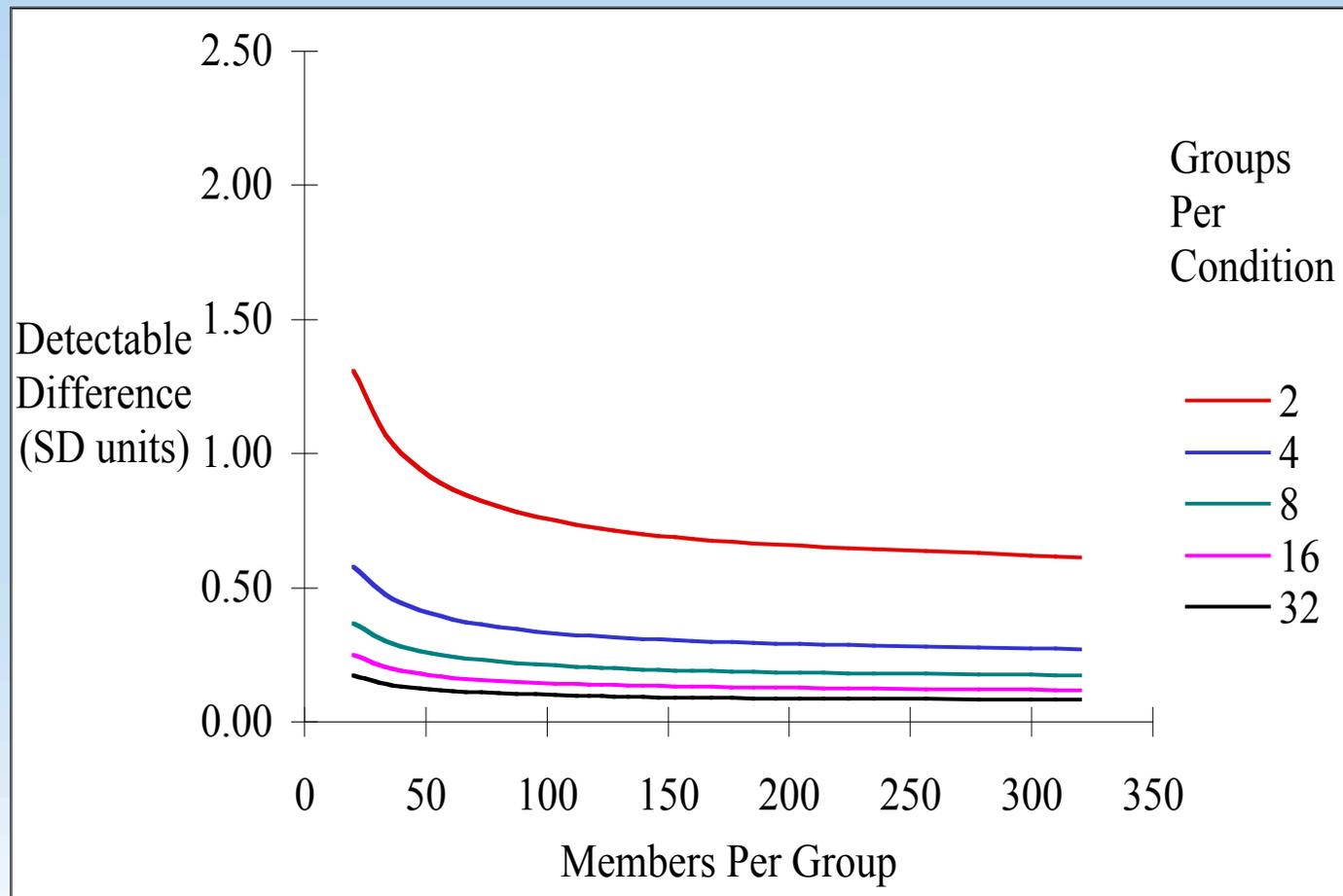
Strategies to Improve Precision

- Increased replication (ICC=0.100)



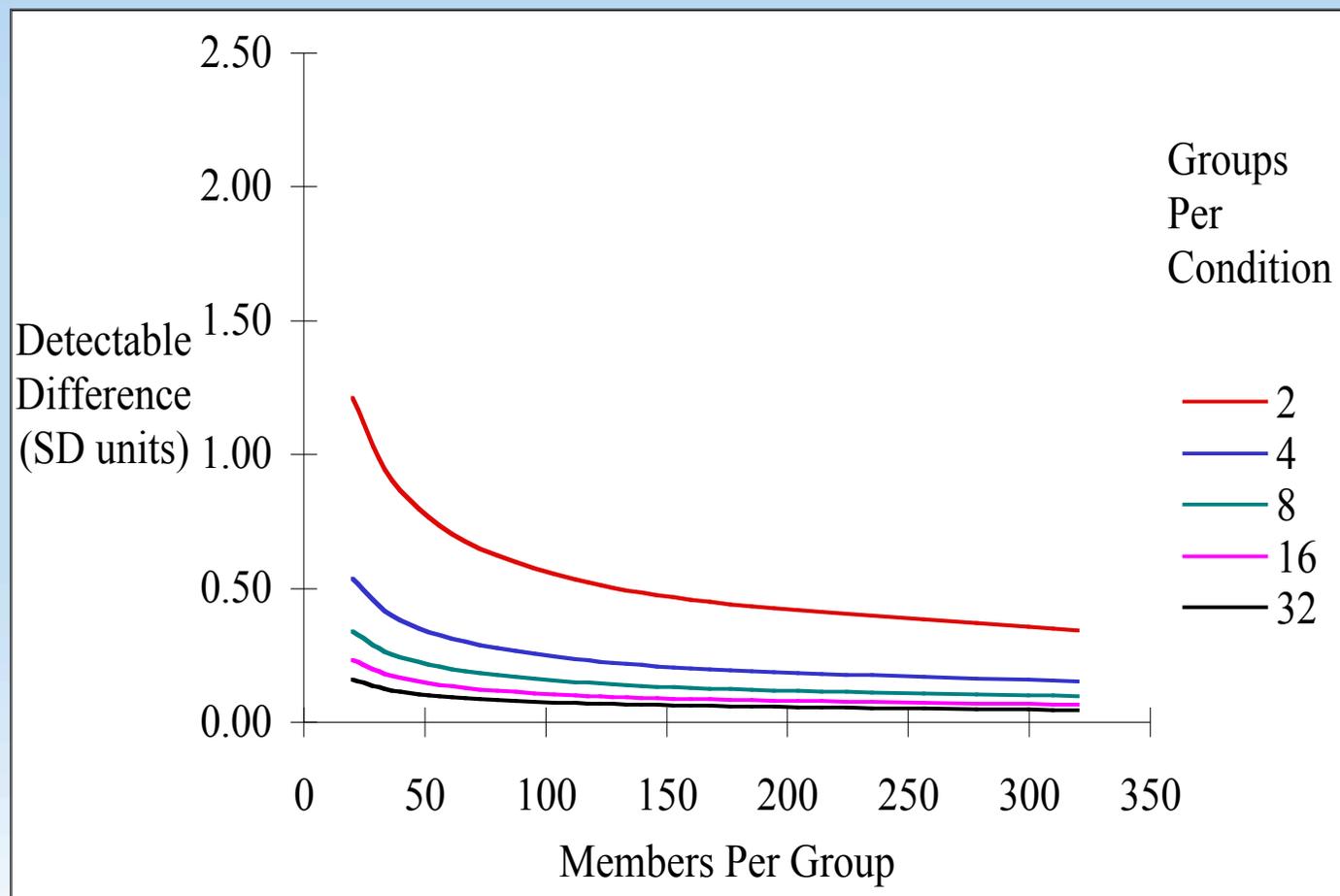
Strategies to Improve Precision

■ Reduced ICC (ICC=0.010)



Strategies to Improve Precision

- The law of diminishing returns (ICC=0.001)



Group-Randomized Trial Designs

- Single factor and factorial designs
- Time as a factor
- Cohort vs. cross-sectional designs
- *A priori* matching and stratification
- *Post hoc* stratification
- Constrained randomization

Single Factor and Factorial Designs

- Most GRTs involve only one treatment factor.
 - Condition
- Most GRTs have only two levels of that treatment factor.
 - Intervention vs. control
- Most GRTs cross Condition with Time.
 - Nested cohort designs
 - Nested cross-sectional designs
- Some GRTs include stratification factors.
 - Multi-center GRTs cross Condition with Field Center.

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

Cross-Sectional and Cohort Designs

- Strengths and weaknesses

Cross-section

in migration 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 an important 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 $m < 100$.

Post Hoc Stratification

- With *a priori* stratification, strata are defined in advance and the units of assignment are randomized to condition x strata cells.
- With *post hoc* stratification, strata are defined *post hoc* and the stratification factor is added to the analysis.
- Common *post hoc* stratification factors include gender, age group, race or ethnic group, etc.
- In designs in which the individual is the unit of assignment, there is no difference between the analysis for *a priori* stratification and the analysis for *post hoc* stratification.
- There is a large difference in GRTs (Murray, 1998).

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 Analytic Strategies 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 Analytic Strategies 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) reported that randomization tests had nominal Type I and II error rates across a variety of 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.

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.

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 been 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 instead of groups is not recommended.

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.

Will Kish's Effective df Help?

- Some have suggested evaluating the intervention effect against effective $df = (\text{individual } df) / DEFF$.
 - 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.

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 $g > 6$.

What About Constrained Randomization?

- Stratification or matching are difficult if there are multiple stratification or matching factors and a limited number of groups to be randomized.
- Constrained randomization has been suggested as a solution (Moulton, 2004).
 - Generate all possible allocations.
 - Identify those that are sufficiently well balanced across conditions
 - Choose one allocation from the constrained set.
 - Use that allocation for the study.

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 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 well used.

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 estimates of the variance components 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).

State of the Science for Analytic Methods in Group-Randomized Trials

- 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.

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.

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).
 - 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.

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.

Power for Group-Randomized Trials

- The usual methods must be adapted to reflect the nested design
 - A good source on power is Chapter 9 in Murray (1998).
 - Other texts include Donner & Klar, 2000; Hayes & Moulton, 2009; Campbell & Walters, 2014.
 - Many papers now report ICCs and show how to plan a GRT.
 - e.g. Murray & Blitstein, 2003; Murray, Catellier et al., 2004; Janega et al., 2004; Hade et al., 2010.
- Power in GRTs is tricky, and investigators are advised to get help from biostatisticians familiar with these methods.
- Power for IRGTs is often even trickier, and the literature is more limited.
 - cf. Pals et al. 2008.

Cornfield's Two Penalties

- Extra variation
 - Condition-level statistic vs. group-level statistic
 - Greater variation in the group-level statistic
 - Reduced power, other factors constant.
- Limited df
 - df based on the number of groups
 - Number of groups in a GRT is often limited
 - Reduced power, other factors constant

Strategies to Reduce Extra Variation

- Effective strategies
 - Sampling methods
 - Random sampling within groups rather than subgroup sampling
 - Timing of measurement
 - Spring surveys rather than fall surveys for school studies (Murray et al., 1994)
 - Spreading surveys over time where there is a high within-day ICC (Murray, Catellier et al, 2004)
 - Regression adjustment for covariates
 - Fixed covariates in non-repeated measures analyses
 - Time-varying covariates in repeated measures analyses

Strategies to Increase df

- Discounted strategies
 - Individual level df (Murray et al., 1996)
 - Kish's effective df (Murray et al., 1996)
 - Subgroup df (Murray et al., 1996)
 - Mixed-model ANOVA/ANCOVA with more than 2 time intervals in the model (Murray et al., 1998)
- Effective strategies
 - Increased replication of groups and members

Sample Size, Detectable Difference and Power

- There are seven steps in any power analysis.
 - Specify the form and magnitude of the intervention effect.
 - Select a test statistic for that effect.
 - Determine the distribution of that statistic under the null.
 - Select the critical values to reflect the desired Type I and II error rates.
 - Develop an expression for the variance of the intervention effect.
 - Gather estimates of the parameters that define that variance.
 - Calculate sample size, detectable difference or power based on those estimates.

Sample Size, Detectable Difference and Power

- Intervention effects are often defined as 1 df contrasts.
 - A t-test is an appropriate test.
 - The shape of the t-distribution is well known.
 - Critical values are easily obtained given the Type I and II error rates.
- Murray (1998) and other sources provide formulae for the variance of the intervention effect.
- The sixth step...
 - Gather estimates of the parameters that define the variance
 - Best done from data that are similar to the data to be collected (similar population, measures, design, and analysis).

Estimating ICC

- From the literature
- From a one-way ANOVA with group as the only fixed effect:

$$ICC_{m:g:c} = \frac{MS_{\text{between}} - MS_{\text{within}}}{MS_{\text{between}} + (m-1)MS_{\text{within}}}$$

Detectable Difference

- The seventh step...
 - Calculate sample size, detectable difference, or power based on those estimates.
 - For a one df contrast between two condition means or mean slopes, the detectable difference in a simple RCT is:

$$\begin{aligned}\hat{\Delta} &= \sqrt{\hat{\sigma}_{\Delta}^2 \left(t_{\text{critical}:\alpha/2} + t_{\text{critical}:\beta} \right)^2} \\ &= \sqrt{2 \left(\frac{\sigma_y^2}{n} \right) \left(t_{\text{critical}:\alpha/2} + t_{\text{critical}:\beta} \right)^2}\end{aligned}$$

Detectable Difference

- The seventh step...
 - Calculate sample size, detectable difference, or power based on those estimates.
 - For a one df contrast between two condition means or mean slopes, the detectable difference in a simple GRT is:

$$\begin{aligned}\hat{\Delta} &= \sqrt{\hat{\sigma}_{\Delta}^2 \left(t_{\text{critical}:\alpha/2} + t_{\text{critical}:\beta} \right)^2} \\ &= \sqrt{2 \left(\frac{\hat{\sigma}_y^2 \left(1 + (m-1) \text{ICC}_{m:g:c} \right)}{mg} \right) \left(t_{\text{critical}:\alpha/2} + t_{\text{critical}:\beta} \right)^2}\end{aligned}$$

Sample Size

- The seventh step...
 - Calculate sample size, detectable difference, or power based on those estimates.
 - For a one df contrast between two condition means or mean slopes, the sample size per condition for a given detectable difference Δ in a simple RCT is:

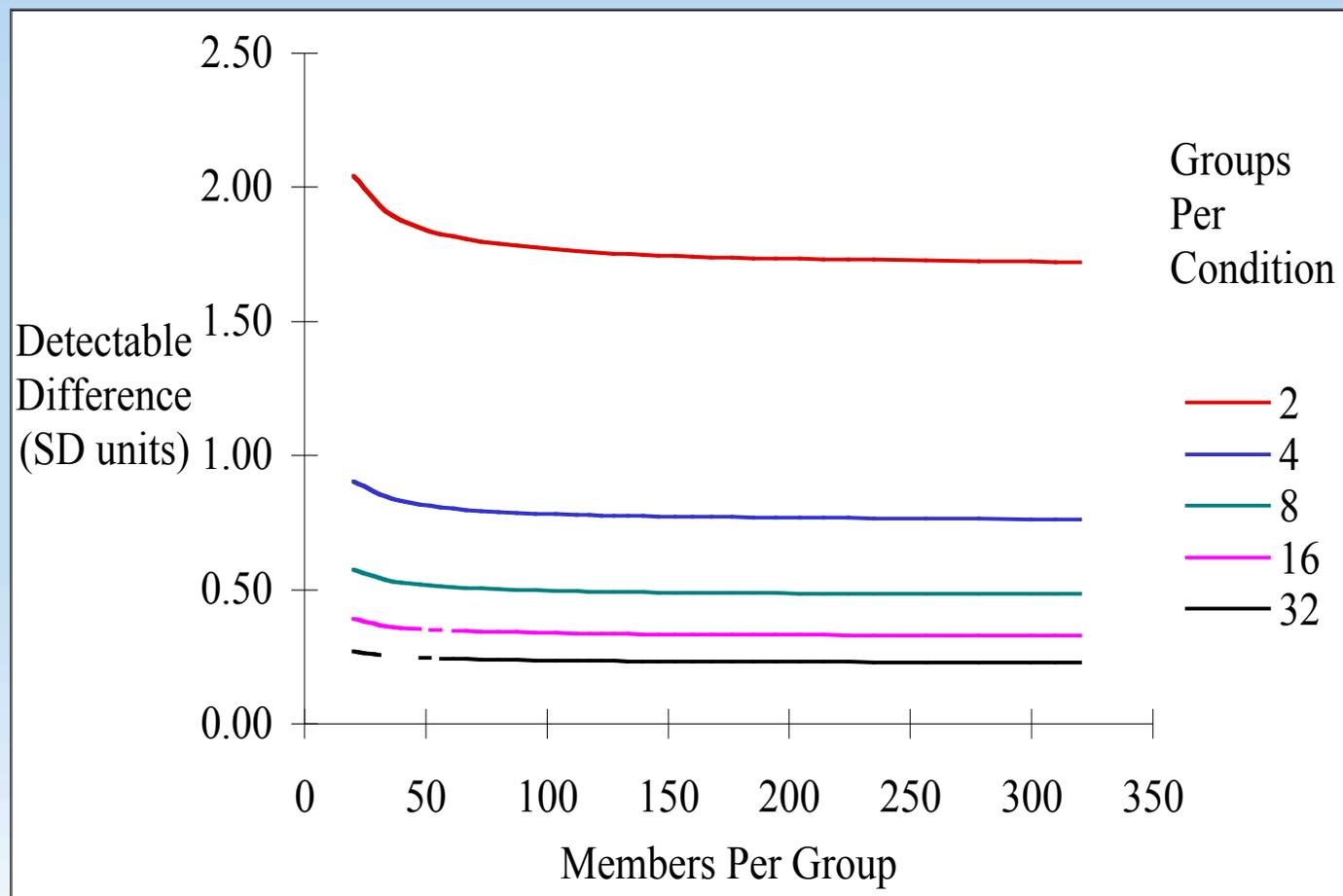
$$n = \frac{2\hat{\sigma}_y^2 \left(t_{\text{critical} : \alpha/2} + t_{\text{critical} : \beta} \right)^2}{\hat{\Delta}^2}$$

- In a simple GRT, this expression becomes:

$$g = \frac{2\hat{\sigma}_y^2 \left(1 + (m-1) \text{ICC}_{m:g:c} \right) \left(t_{\text{critical} : \alpha/2} + t_{\text{critical} : \beta} \right)^2}{m \hat{\Delta}^2}$$

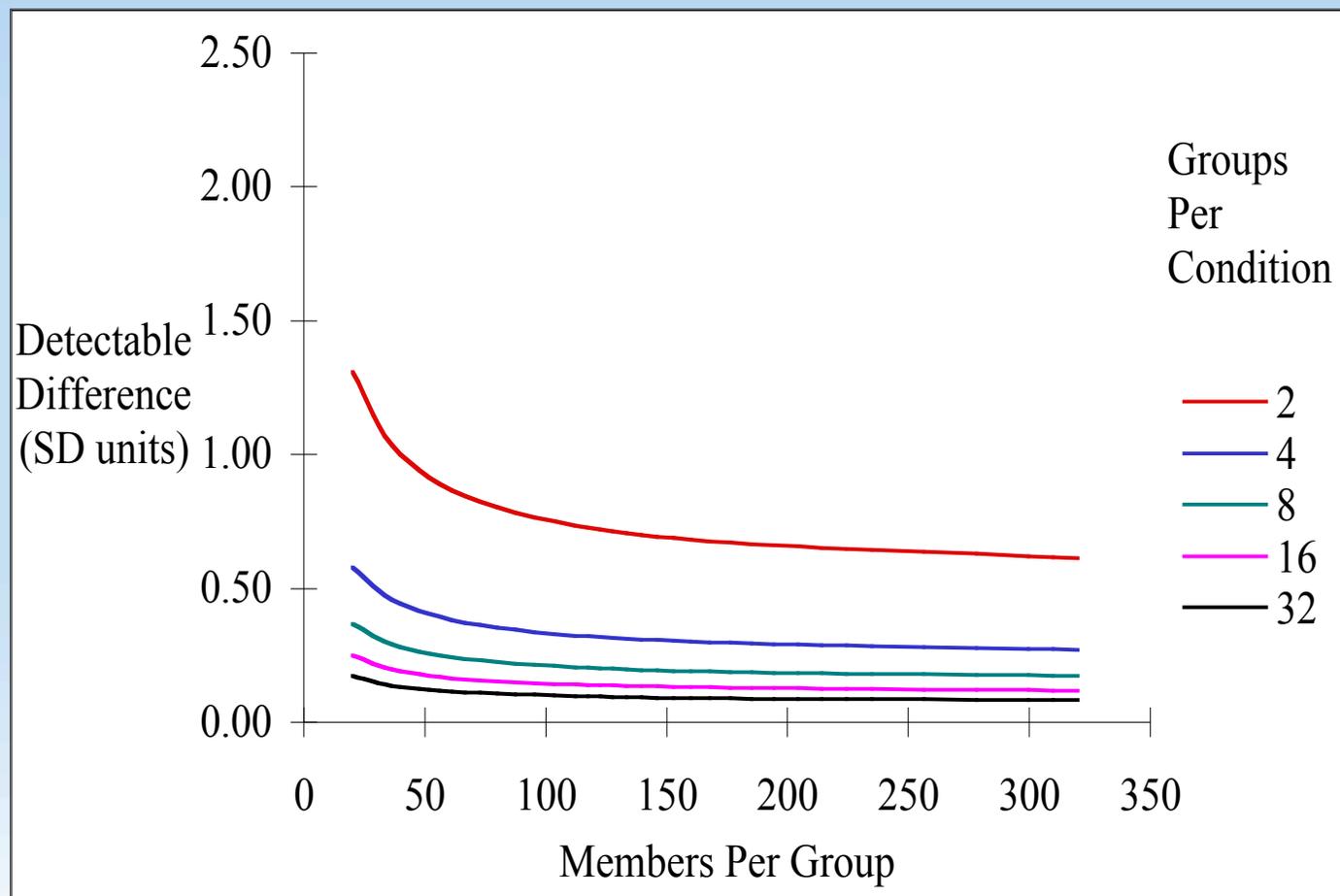
Detectable Difference

- The most influential factors are the ICC and g . (ICC=0.100)



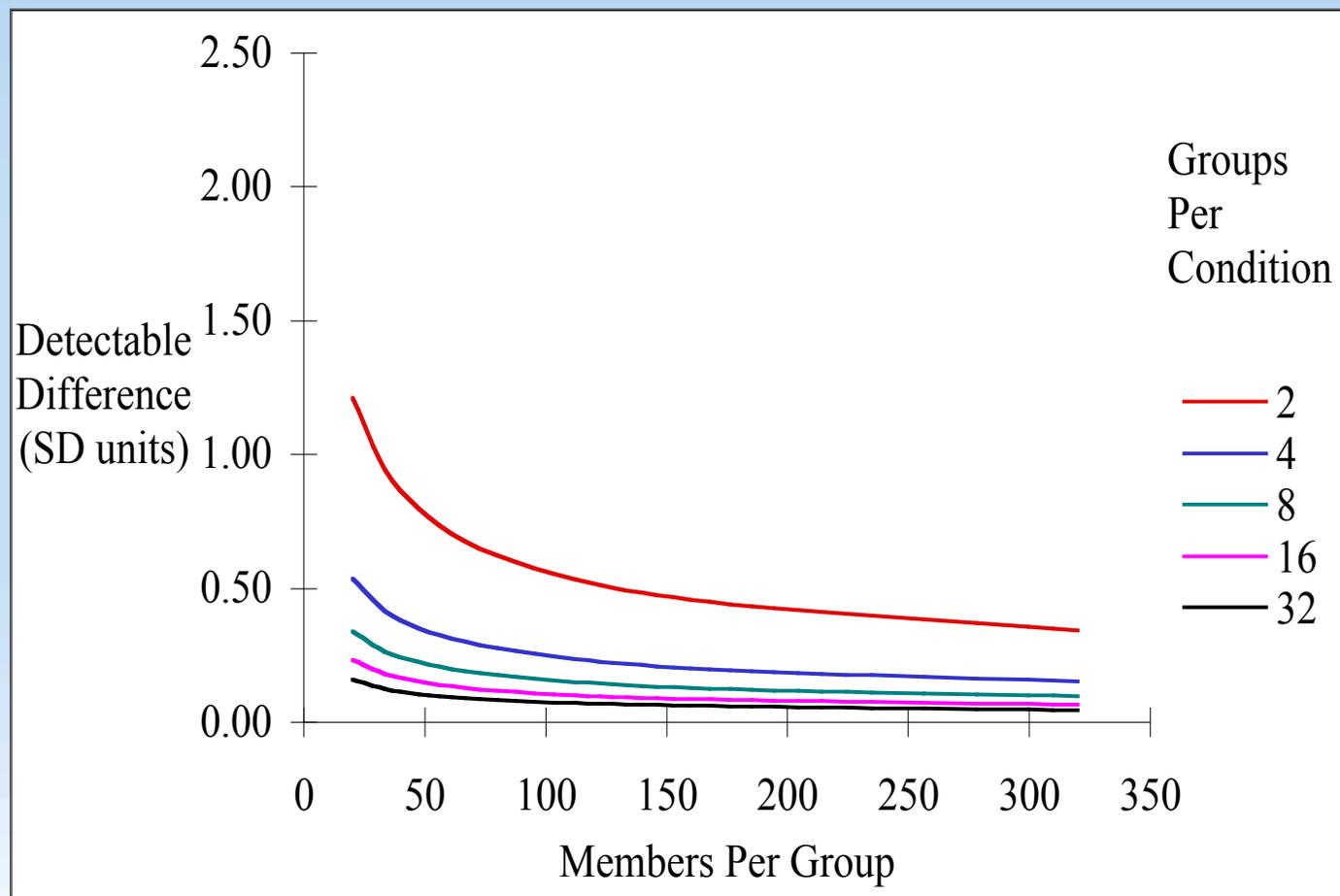
Detectable Difference

- The most influential factors are the ICC and g. (ICC=0.010)



Detectable Difference

- The most influential factors are the ICC and g . (ICC=0.001)



What About Alternative Designs?

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

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

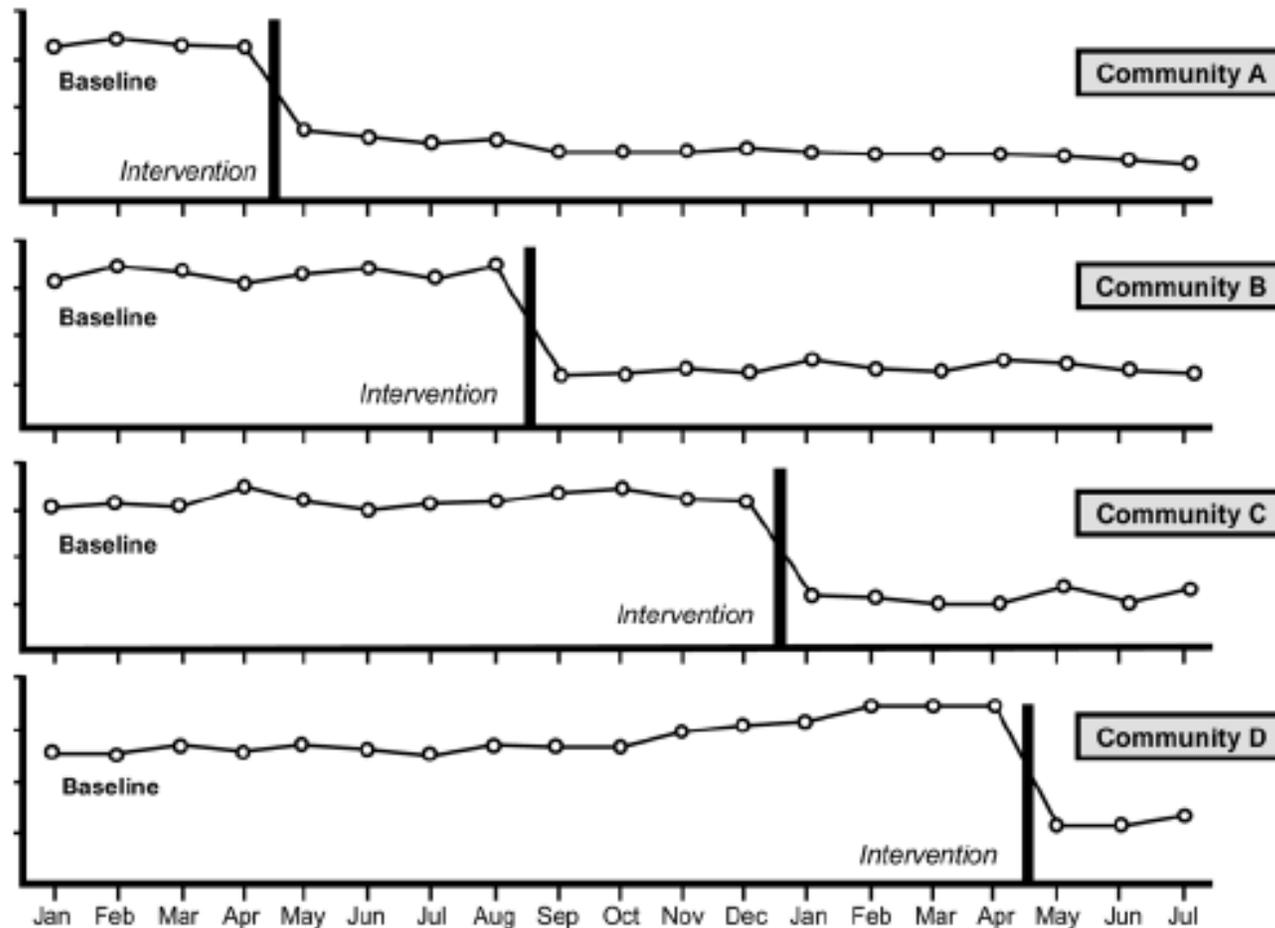


Figure 1. Hypothetical example of a multiple baseline design used to assess behavior change following an intervention in four communities.

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.
- cf. Rhoda et al., 2011.

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

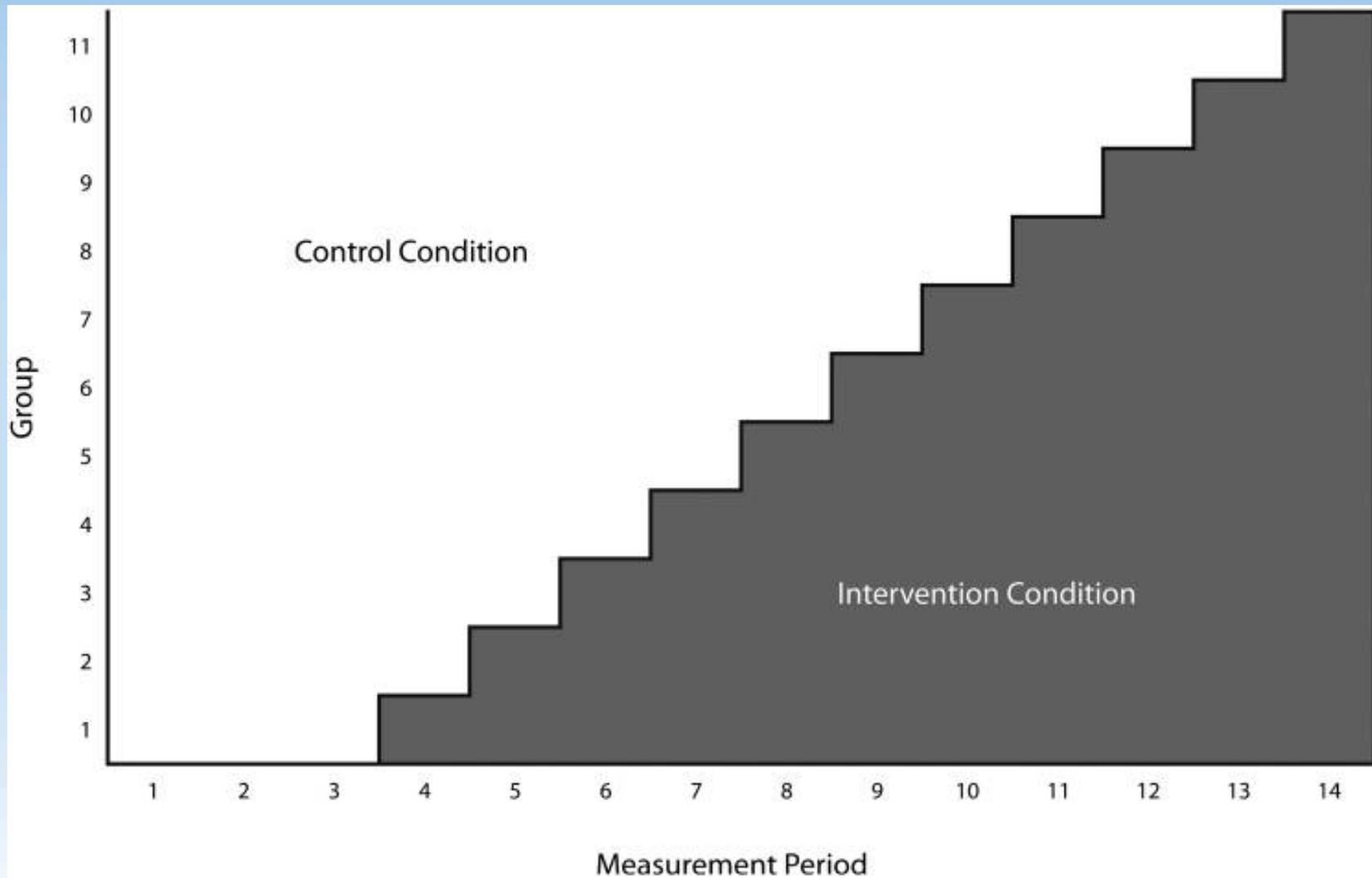
- QEs have all the features of experiments except randomization.
 - Causal inference requires elimination of plausible alternatives.
- If groups are assigned and members are observed, analysis and power issues are the same as in GRTs.
- Useful when randomization is not possible.
 - Can provide experience 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.
- cf. Shadish et al., 2000.

Stepped-Wedge Designs

- Sometimes called Dynamic Wait-List Designs
- Combine the features of multiple baseline designs and GRTs.
 - Measurement is frequent and on the same schedule in all groups.
 - Time is divided into intervals.
 - Groups selected at random for the intervention in each interval.
 - By the end of the study, all the groups have the intervention.
- Both *Trials* (2015) and the *Journal of Clinical Epidemiology* (2013) recently published issues focused on the design and analysis of stepped wedge designs.

- Cf. also Hughes et al. (2015)

Stepped Wedge Design



Stepped Wedge Design

- The analysis estimates a weighted average intervention effect across the intervals.
 - Assumes that the intervention effect is rapid and lasting.
 - Not very sensitive to intervention effects that develop gradually or fade over time.
- These designs can be more efficient but usually take longer to complete and cost more than the standard GRT.
- cf. Rhoda et al, 2011.

Regression Discontinuity Designs

- Groups or 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., 2014, 2015; O'Keefe et al., 2014).

Regression Discontinuity Design

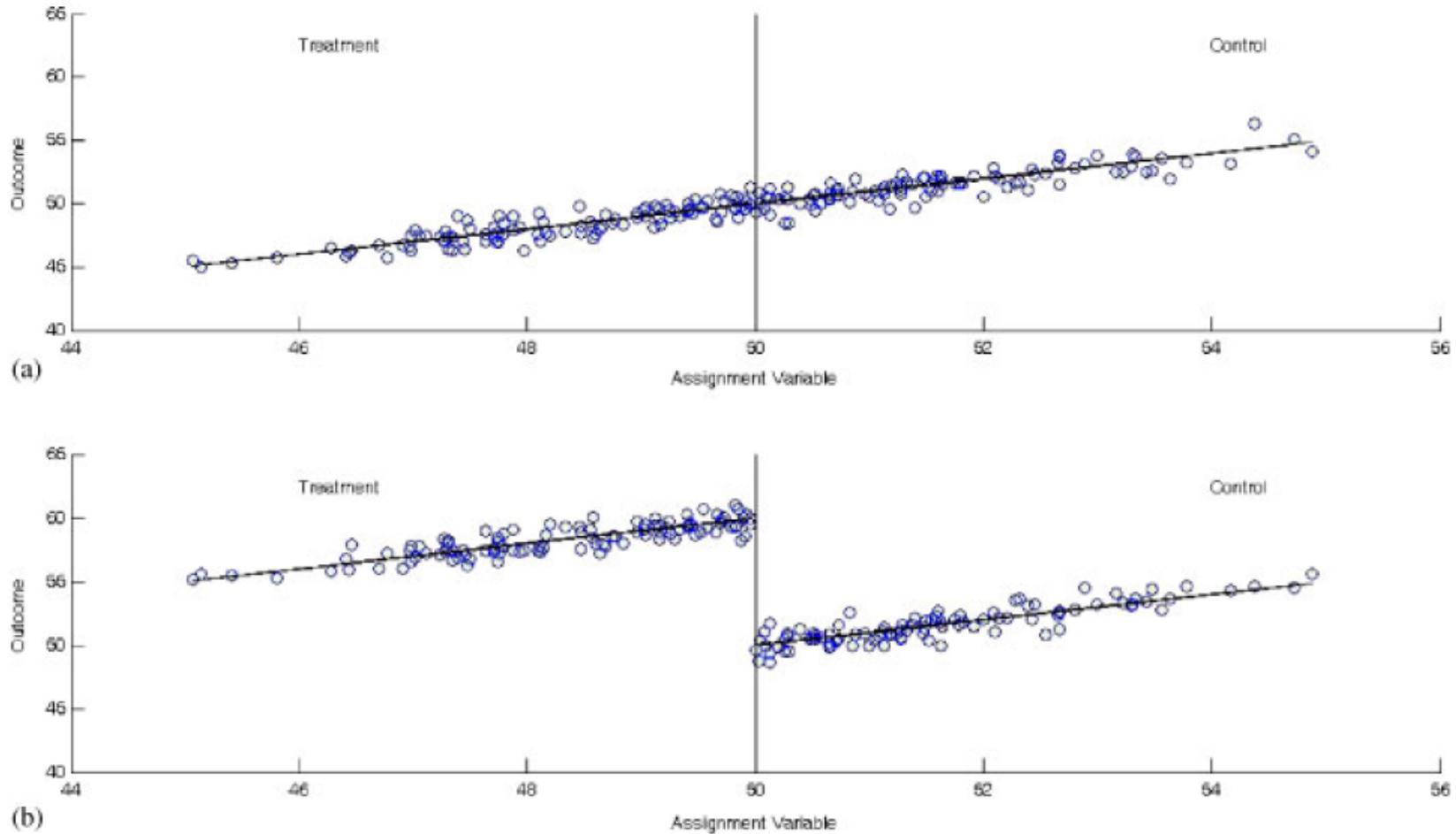


Figure 1. Hypothetical regression discontinuity experiments: (a) ineffective treatment and (b) effective treatment.

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.
 - Sample size requirements are usually doubled.
- cf. Pennell et al., 2011.

Summary

- A 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
- GRTs provide better 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 GRTs have been proposed.
 - Multiple baseline designs
 - Time series designs
 - Quasi-experimental designs
 - Dynamic wait-list or stepped-wedge designs
 - Regression discontinuity designs
- Under the right conditions, these alternatives can provide good 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 and may take longer to complete
 - Quasi-experimental designs, stepped wedge, regression discontinuity

Examples

- Group-randomized trials: Health Care Systems Collaboratory
 - 9 pragmatic trials conducted in collaboration with health care systems, funded as UH2/UH3 trials by a variety of NIH ICs.
 - 8 are group-randomized trials (GRT)
 - Hospital acquired infections
 - **CRC screening (STOP CRC)**
 - Healthcare utilization in spinal injuries
 - Chronic pain management
 - Mortality in dialysis patients
 - Management of PTSD in trauma patients
 - Advanced care planning in nursing homes
 - Management of multiple chronic conditions

Strategies and Opportunities to STOP CRC in Priority Populations

- Key personnel
 - PI: Gloria Coronado, PhD
 - Statistician: Bill Vollmer, PhD
 - Institution: Kaiser Permanente Center for Health Research
- Primary objective
 - Test the effectiveness of automated EMR-driven strategies to raise CRC screening rates in safety-net clinics
- Primary outcome
 - Proportion of targeted patients who complete FIT kit during first year of intervention.

STOP CRC Design

- Group-randomized trial
 - 26 federally qualified health clinics
 - Affiliated with 8 larger administrative networks
 - Clinic-level randomization stratified by network
 - EMR used to drive system-level intervention
 - Control clinics roll out intervention in year 2
 - Consent waived for this minimal risk study

- Illustrates *a priori* stratification in a GRT, with clinic as the unit of assignment and a delayed-treatment control condition.

STOP CRC Analytic Approach

- Weighted logistic regression accounting for clustering at clinic level and adjusting for selected individual and clinic level covariates.
 - cf. Coronado et al., 2014 for details on the design and analytic plan.
-
- Illustrates a mixed-model ANCOVA approach adapted to a dichotomous primary outcome.

STOP CRC

■ Challenges

- Overlap of year 1 measurement window and year 2 intervention rollout for control clinics
- Use of real-time EMR tools that may be discordant with static randomization tables
- Implementation delays and ACA rollout

- These challenges threatened the validity of the primary analysis

STOP CRC

■ Solutions

- Delayed rollout of intervention for control clinics in year 2 to deal with the overlap problem.
 - Formulated a number of sensitivity analyses to try to overcome impact of lags in startup and hence give a more accurate estimate of true intervention impact.
 - Include a stepped wedge framework in which data from both years 1 and 2, as well as year prior to randomization, are used to estimate separate startup effects in year 1 of intervention and steady state effects in year 2 of intervention.
-
- Adaptations required during planning year to accommodate real world complexities.

Examples

- Group-randomized trials: Health Care Systems Collaboratory
 - 9 pragmatic trials conducted in collaboration with health care systems, funded as UH2/UH3 trials by a variety of NIH ICs.
 - 8 are group-randomized trials (GRT)
 - Hospital acquired infections
 - CRC screening
 - Healthcare utilization in spinal injuries
 - **Chronic pain management (PACT)**
 - Mortality in dialysis patients
 - Management of PTSD in trauma patients
 - Advanced care planning in nursing homes
 - Management of multiple chronic conditions

Collaborative Care for Chronic Pain in Primary Care

PACT

- Key personnel
 - PI: Lynn DeBar, PhD, MPH
 - Statistician: Bill Vollmer, PhD
 - Institution: Kaiser Permanente Center for Health Research
- Primary objective
 - Test whether an integrative pain management program embedded within primary care: decreases pain, opioid use, and healthcare utilization; and improves function for patients with complex chronic pain
- Primary outcome
 - Trajectory of change in self-reported pain scores over the first six months of intervention

PACT Design

- Stratified group-randomized trial
 - Strata are three regions of the Kaiser Permanente Health Plan
 - Physicians are unit of randomization
 - EMR screen to identify potentially eligible patients
 - Vet list with PCPs
 - Verbal consent obtained from patients prior to randomization

- Illustrates stratified group-randomized trial with physician as the unit of assignment.

PACT Analytic Approach

- Two-stage analysis
 - Compute slopes for individual pain score trajectories
 - Analyze slopes using mixed model ANCOVA adjusting for selected individual and cluster level variables, including baseline pain score
 - Cf. DeBar et al., 2012 for details on the rationale for this approach.
- Illustrates two-stage analysis with regression adjustment for covariates.

PACT

■ Challenges

- Weaving a complex, multi-modal intervention into fabric of usual care
- Everyone doing things/creating partnerships never done before:
 - Redeploying/hiring clinical staff for intervention roles not well-aligned with existing health plan structure or traditional scope of practice
 - Expanding use of EHR
 - Creating scalable training model with attention to fidelity and cost/resources
 - Sharing costs and building infrastructure processes
 - IRBs uneasy relinquishing tight research constraint.
- Pragmatic trials are not easy, especially working in new systems with new methods for data collection and intervention delivery.

PACT

■ Solutions

- Had to adapt the intervention structure to accommodate clinical work flow and stakeholder input.
 - Had to redefine some clusters by grouping PCPs due to smaller than expected number of consenting patients for some PCPs.
 - Delayed startup in some regions until systems could be put in place to properly implement the intervention.
 - Shifted projected N between regions to reflect what was possible.
 - Team has been forced to devote a larger proportion of their effort than anticipated to solve implementation issues.
- Pragmatic trials are not easy, especially working in new systems with new methods for data collection and intervention delivery.

Examples

- Group-randomized trials: Health Care Systems Collaboratory
 - 9 pragmatic trials conducted in collaboration with health care systems, funded as UH2/UH3 trials by a variety of NIH ICs.
 - 8 are group-randomized trials (GRT)
 - Hospital acquired infections
 - CRC screening
 - Healthcare utilization in spinal injuries
 - Chronic pain management
 - Mortality in dialysis patients
 - **Management of PTSD in trauma patients (TSOS)**
 - Advanced care planning in nursing homes
 - Management of multiple chronic conditions

Trauma Survivors Outcomes & Support TSOS

- Key personnel
 - PI: Douglas Zatzick, MD
 - Statistician: Patrick Heagerty, PhD
 - Joan Russo, Bryan Comstock, Jin Wang
 - Institution: University of Washington
- Primary objective
 - Explore intervention effect in patients with pre-injury chronic medical conditions
- Primary outcome
 - PTSD symptoms

TSOS Design

- Stepped wedge design
 - 24 US Level I trauma centers randomized to 4 waves
 - 960 patients with PTSD (40 patients/trauma center)
 - All co-morbidities included
 - All trauma centers recruit both control and intervention patients
 - All trauma centers begin recruiting controls
 - Data collected at baseline, 3, 6, and 12 months
 - Intervention “turned on” at each trauma center per design
 - Implementation advantage: all trauma centers trained
 - Design adds analytic complexity

- Illustrates stepped wedge design.

TSOS Analytic Approach

- Intervention vs. Control Comparisons
 - PTSD (Primary)
 - Alcohol
 - Depression
- Subgroup Analyses
 - Pre-injury Medical Conditions (ICD)
 - Traumatic brain injury (ICD)
- cf. Hughes et al. 2015 for a discussion of some of the analysis issues in stepped wedge designs.
- Illustrates mixed effect regression approach with adjustment for covariates.

TSOS

- Challenges Raised by 24 site Design
 - Site Variability
 - Sites vary in rates of violent injury (↑PTSD with ↑violence)
 - Sites vary in other characteristics (e.g., admission volumes)
 - Implementation challenge
 - In consideration of American College of Surgeons mandate for PTSD screening and intervention, all sites want intervention training

TSOS

- Solution: Stepped Wedge Design
 - Site Variability: Each site contributes control & intervention patients
 - Implementation challenge: All sites receive intervention training

Examples

- **Group-randomized trials: Health Care Systems Collaboratory**
 - 9 pragmatic trials conducted in collaboration with health care systems, funded as UH2/UH3 trials by a variety of NIH ICs.
 - 8 are group-randomized trials (GRT)
 - Hospital acquired infections
 - CRC screening
 - Healthcare utilization in spinal injuries
 - Chronic pain management
 - Mortality in dialysis patients
 - Management of PTSD in trauma patients
 - Advanced care planning in nursing homes
 - **Management of multiple chronic conditions (PIECES)**

Improving Chronic Disease Management with Pieces™

- Key personnel
 - PI: Miguel Vazquez, MD
 - Biostatisticians: Chul Ahn, PhD and Song Zhang, PhD
 - Institution: University of Texas Southwestern Medical Center
- Primary objective
 - To evaluate the management of patients with CKD, diabetes, and hypertension with a clinician support model enhanced by technology support (Pieces™) compared with standard of care.
- Primary outcome
 - 1-year all cause hospitalization

Pieces™ Design

- Stratified group-randomized trial
 - Four healthcare systems with 249 clinics and >35,000 patients available.
 - Within each healthcare system, clinics or practice sites will be randomized to either Pieces™ or standard care group.
 - Every patient assigned to a given clinic or practice site will receive the intervention to which the clinic or practice site was randomized.

- Illustrates stratified group-randomized trial with clinic or practice site as the unit of assignment.

Pieces™ Analytic Approach

- Primary analysis
 - The generalized Mantel-Haenszel testing procedure (Donner 1992) will be applied to detect any difference in hospitalization rate between Pieces™ and standard care.
- Secondary analysis
 - Mixed logistic regression to assess intervention effect on hospitalization rate controlling for clustering and patient, clinician, and clinic factors.
 - Cox models to assess the intervention effect on time to hospitalization with frailty to control for clustering.
- Illustrates non-parametric approach to primary analysis and model-based approach to secondary analysis.

Pieces™

■ Challenges

- Getting informed consent waivers.
 - Resolving heavy work loads among participating centers.
 - Streamlining clinical workflows for each site
 - Competing priorities for IT build
 - Slow approval process at one of the study healthcare systems
 - Training of PCPs and staff at each clinic site
-
- Such logistical issues are common in pragmatic trials in the health care setting

Pieces™

- Solutions
 - The team is currently addressing these logistical issues.

A Review of Recent Practices in GRTs

- Murray, D. M., S. P. Pals, et al. (2008). Design and analysis of group-randomized trials in cancer: a review of current practices. *Journal of the National Cancer Institute*, 100(7): 483-491.

A Review of Recent Practices in GRTs

Previous Reviews of the GRT Literature

- The first review was published by Donner et al. in 1990.
 - Only 19% took the ICC into account in the sample size calculations.
 - Only 50% took the ICC into account in the analysis.
- A review by Simpson et al. in 1995 reported little progress.
 - Only 19% took the ICC into account in the sample size calculations.
 - Only 57% took the ICC into account in the analysis.
- A review by Varnell et al. in 2004 reported no progress, though the standards were higher than in previous reviews.
 - Only 15% took the ICC into account in the sample size calculations.
 - Only 54% always took the ICC into account in the analysis.
- We were interested in whether the situation had improved.

A Review of Recent Practices in GRTs Procedures

- Systematic review of cancer-related studies published 2002-2006.
 - Medline and PubMed search.
 - Studies had as their primary outcome cancer risk factors, cancer morbidity, or cancer mortality.
 - Studies used randomization to assign identifiable social groups to study conditions, with observations taken on members of those groups to assess the impact of an intervention.
 - Where the paper referred to an earlier "design paper", we also reviewed that paper.
 - Each reviewer independently assessed the article on items related to design, sample size estimation, and analysis.
 - The reviewers discussed each paper as a group and any disagreements were resolved in discussion.

A Review of Recent Practices in GRTs

Findings

- 92 possible group-randomized trials in 45 journals.
- 75 articles from 41 journals that met the inclusion criteria.
- 20 background "design" papers.
- 20% in the *Preventive Medicine*
- 7% in *American Journal of Public Health*
- No more than 4% in any other single journal
- 15.0 GRT papers per year (2002-06) vs 11.6 per year (1998-2002) in Varnell et al. and 5.3 per year (1990-93) in Simpson et al.

Table 1. Analytic methods frequently used in group-randomized trials and the conditions under which their use is appropriate

Method	Appropriate application in group-randomized trials
Mixed-model methods	
Repeated measures ANOVA/ANCOVA*	Outcomes for which there are one or two time points, variation in means or rates at the condition level is assessed against variation in means or rates at the group level, and df are based on the number of groups
Random coefficient approach	Outcomes for which there are three or more time points, variation in slopes and intercepts at the condition level is assessed against variation in slopes and intercepts at the group level, and df are based on the number of groups
Generalized estimating equations	
With correction for limited df	Trials where there are fewer than 40 df available for the test of the intervention effect
With no correction	Trials where there are 40 or more df available for the test of the intervention effect

* ANOVA = analysis of variance; ANCOVA = analysis of covariance.

Table 1. Analytic methods frequently used in group-randomized trials and the conditions under which their use is appropriate

Method	Appropriate application in group-randomized trials
Two-stage methods	Calculation of group means or other summary statistic in the first stage and analysis of variation in those summary statistics at the condition level in the second stage, df based on the number of groups
Post hoc correction based on external estimates of intraclass correlation	To data for which there are valid external estimates of intraclass correlation
Analysis at individual level, ignoring group-level intraclass correlation	None for group-randomized trials
Analysis at subgroup level, ignoring group-level intraclass correlation	None for group-randomized trials

* ANOVA = analysis of variance; ANCOVA = analysis of covariance.

Table 3. Characteristics of 75 articles that reported the results of group-randomized trials from cancer research in selected peer-reviewed journals during the period 2002 – 2006, inclusive *

Characteristic	No.	%
No. of study conditions		
2	66	88
3	5	6.7
≥ 4	4	5.3
Design		
Cohort	50	67
Cross-sectional	15	20
Combination of cohort and cross-sectional	10	13
Matching or stratification in design		
Matching only	15	20
Stratification only	26	35
Matching and stratification	4	5
Randomization without matching or stratification	30	40

* Percentages within subsections of may not add to the subsection total because the categories were not always mutually exclusive.

Table 3. Characteristics of 75 articles that reported the results of group-randomized trials from cancer research in selected peer-reviewed journals during the period 2002 – 2006, inclusive *

Characteristic	No.	%
Type of group		
Churches	4	5.3
Communities, neighborhoods, or community groups	14	19
Housing projects or apartment buildings	3	4.0
Physicians or provider groups	28	37
Schools or colleges	15	20
Worksites	11	15
No. of groups per condition		
1	4	5.3
2 – 5	21	28
6 – 8	10	13
9 – 12	9	12
13 – 25	17	23
>25	12	16
Not reported	2	2.7

* Percentages within subsections of may not add to the subsection total because the categories were not always mutually exclusive.

Table 3. Characteristics of 75 articles that reported the results of group-randomized trials from cancer research in selected peer-reviewed journals during the period 2002 – 2006, inclusive *

Characteristic	No.	%
No. of members per group		
<10	14	19
10 – 50	22	29
51 – 100	15	20
>100	17	23
>1 type of member	5	6.7
Not reported	2	2.7
No. of time points		
1	41	54
2	22	29
3 – 9	4	5.4
No. of time points varies within study	7	9.3
Unknown or continuous	1	1.3

* Percentages within subsections of may not add to the subsection total because the categories were not always mutually exclusive.

Table 3. Characteristics of 75 articles that reported the results of group-randomized trials from cancer research in selected peer-reviewed journals during the period 2002 – 2006, inclusive *

Characteristic	No.	%
Focus of study		
Primary prevention	39	52
Secondary prevention	24	32
Tertiary prevention	5	6.7
Combination	5	6.7
None of the above	2	2.7
Target population		
Individuals with no personal history of the target cancer	14	19
Cancer survivors during primary treatment	3	4.0
Cancer survivors after primary treatment	1	1.3
Unknown or mixed cancer survivorship	49	65
Caregivers	3	4.0
Combination	5	6.7

* Percentages within subsections of may not add to the subsection total because the categories were not always mutually exclusive.

Table 3. Characteristics of 75 articles that reported the results of group-randomized trials from cancer research in selected peer-reviewed journals during the period 2002 – 2006, inclusive *

Characteristic	No.	%
Primary outcome variables		
Alcohol use	2	2.7
Delivery of health services	7	9.3
Dietary variables	11	15
Incidence of cancer	2	2.7
Knowledge of cancer or attitudes regarding cancer	9	12
Mortality from cancer	2	2.7
Physical activity	7	9.3
Screening	24	32
Sun protection	6	8.0
Tobacco use	15	20
Other	3	4.9

* Percentages within subsections of may not add to the subsection total because the categories were not always mutually exclusive.

Table 4 . Distribution of analytic methods in 75 articles that reported the results of group-randomized trials from cancer research published in selected peer-reviewed journals during the period 2002 – 2006 inclusive *

Criteria	No.	%
Articles reporting only appropriate methods	34	45
Mixed-model methods with adjustment for baseline or other covariates	15	20
Mixed-model repeated measures with two time points	9	12
Random coefficient model with more than two time points	1	1.3
Generalized estimating equations with ≥ 40 groups	8	11
Two-stage analysis (analysis of group means or other summary statistics)	4	5.3

* Percentages within subsections of may not add to the subsection total

Table 4 . Distribution of analytic methods in 75 articles that reported the results of group-randomized trials from cancer research published in selected peer-reviewed journals during the period 2002 – 2006 inclusive *

Criteria	No.	%
Articles reporting some appropriate and some inappropriate methods	6	8.0
Appropriate methods		
Mixed-model methods with adjustment for baseline or other covariates	4	5.3
Mixed-model repeated measures with two time points	1	1.3
Random coefficient model with more than two time points	0	0.0
Generalized estimating equations with ≥ 40 groups	1	1.3
Two-stage analysis (analysis of group means or other summary statistics)	0	0.0

* Percentages within subsections of may not add to the subsection total

Table 4 . Distribution of analytic methods in 75 articles that reported the results of group-randomized trials from cancer research published in selected peer-reviewed journals during the period 2002 – 2006 inclusive *

Criteria	No.	%
Articles reporting some appropriate and some inappropriate methods	6	8.0
Inappropriate methods		
Analysis at an individual level, ignoring group-level intraclass correlation	5	6.7
Analysis at a subgroup level, ignoring group-level intraclass correlation	0	0.0
Generalized estimating equations or other asymptotically robust method with <40 groups	0	0.0
Mixed-model repeated measures with more than two time points	0	0.0
Other	1	1.3

* Percentages within subsections of may not add to the subsection total

Table 4 . Distribution of analytic methods in 75 articles that reported the results of group-randomized trials from cancer research published in selected peer-reviewed journals during the period 2002 – 2006 inclusive *

Criteria	No.	%
Articles reporting only inappropriate methods	26	35
Analysis at an individual level, ignoring group-level intraclass correlation	17	23
Analysis at a subgroup level, ignoring group-level intraclass correlation	4	5.3
Analysis with group as a fixed effect	0	0.0
Mixed-model repeated measures with more than two time points	2	2.7
Generalized estimating equations with <40 groups	4	5.3
Not enough information provided	9	12

* Percentages within subsections of may not add to the subsection total

A Review of Recent Practices in GRTs

Discussion

- Our results were similar to those in earlier reviews.
 - 45% of the articles reported only analyses judged to be appropriate.
 - cf. 50-57% in earlier reviews.
 - 35% reported only analyses deemed inappropriate.
 - 8% reported a mix of appropriate and inappropriate analyses.
- Crespi et al. (2011) published a more recent review, but limited to cancer screening studies.
 - 1995-99 55%
 - 2000-02 82%
 - 2003-06 92%
 - 2007-10 55%
- Both reviews suggest there is much room for improvement.

A Review of Recent Practices in IRGTs

- Pals, S. P., Murray, D. M., et al. (2008). Individually randomized group treatment studies: are the most frequently used analytic models misleading? *American Journal of Public Health*, available online.

A Review of Recent Practices in IRGTs Procedures

- There were no prior systematic reviews of IRGT trials.
- We manually searched six journals for the period 2002-06.
 - American Journal of Public Health
 - Preventive Medicine
 - Health Psychology
 - Obesity Research
 - Addictive Behaviors
 - AIDS and Behavior
- Procedures parallel to those used for the GRT review
- Criteria for sample size and analysis methods parallel to those used for the GRT review
- 34 eligible articles

TABLE 2- Characteristics of the Studies Described in 34 Articles Reviewed 2002-2006

Study characteristics	Number of articles	%
Journal		
American Journal of Public Health	4	11.8
Preventive Medicine	6	17.6
Health Psychology	8	23.5
Obesity	7	20.6
Addictive Behaviors	7	20.6
AIDS and Behavior	2	5.9
Year of publication		
2002	5	14.7
2003	6	17.6
2004	6	17.6
2005	5	14.7
2006	12	35.3

TABLE 2- Characteristics of the Studies Described in 34 Articles Reviewed 2002-2006

Study characteristics	Number of articles	%
Number of study conditions ^a		
Two	23	67.6
Three	8	23.5
Four	3	8.8
Number of group treatment conditions ^b		
One	11	32.3
Two	17	50
Three	4	11.8
Four	2	5.9
Baseline sample size		
<100	15	44.1
100-<200	9	26.5
200-<300	4	11.8
>300	6	17.6

TABLE 2- Characteristics of the Studies Described in 34 Articles Reviewed 2002-2006

Study characteristics	Number of articles	%
Target population		
Adults or adolescents with mental health issues	3	8.8
Overweight or obese children	2	5.9
Overweight or obese adults	9	26.5
Adults with cardiovascular risk factors other than weight	3	8.8
Cancer patients	2	5.9
College or University students	2	5.9
HIV-positive adults	3	8.8
Smokers or substance abusers	7	20.6
Other	3	8.8

TABLE 2- Characteristics of the Studies Described in 34 Articles Reviewed 2002-2006

Study characteristics	Number of articles	%
Primary Outcome Variable ^c		
Weight, BMI, Body Fat percentage or Dietary Variables	13	38.2
Physical activity/ physical fitness variables	5	14.7
Smoking or substance use variables	7	20.6
Mental health variables	6	17.6
Sex behavior variables	6	17.6
Treatment retention	2	5.9
Medication adherence	2	5.9
Other variables	7	20.6

TABLE 3- Results of the Review of Sample Size Calculations and Analytic Methods in 34 Articles Reviewed, 2002-2006

Study characteristics	Number of articles	%
Sample size calculations		
Authors reported sample size calculations at individual level	6	17.6
Authors stated power calculations performed, but no detail	1	2.9
No mention of sample size calculation	25	73.5
Authors claimed sample size accounted for ICC, but no detail	1	2.9
Other	1	2.9
Any significant results reported		
Yes	27	79.4
No	7	20.6

TABLE 3- Results of the Review of Sample Size Calculations and Analytic Methods in 34 Articles Reviewed, 2002-2006

Study characteristics	Number of articles	%
Analytic approaches ^a		
Analysis at an individual level, ignoring group entirely	32	94.1
Mixed-model approach with baseline as covariate	2	5.9
Structural equation modeling	1	2.9
Appropriateness of analytic methods		
All analytic methods appropriate	1	2.9
No analytic methods appropriate	32	94.1
Not enough information	1	2.9

A Review of Recent Practices in IRGTs

Discussion

- Warnings have appeared in the literature for at least 30 years regarding the development of intraclass correlation in IRGTs.
- Even so, the literature on the design and analysis of IRGTs is limited.
- The use of inappropriate design and analytic methods is pervasive for IRGTs.
- The picture is similar to what GRTs looked like in the mid 1970s.
- Hopefully, the pattern will improve with time.

Summary

- Pragmatic trials were introduced in 1967 as an approach to help clinicians choose among options for care.
- Pragmatic trials are becoming common in health care settings, where group- or cluster-randomized designs are often used to avoid contamination across study conditions.
- Clinical investigators can learn from the methods developed in public health for group-randomized trials and for alternative designs.
- These methods are well suited to the multilevel interventions that are being developed to treat many complex health issues.
- The research team also will need experience in the systems being considered for the trial, including measurement, operations, and informed consent.

References

■ Primary References

- Murray, D.M. Design and Analysis of Group-Randomized Trials. New York: Oxford University Press, 1998.

■ Secondary References

- Anderson ML, Califf RM, Sugarman J, participants in the NIHC SRCRTW. Ethical and regulatory issues of pragmatic cluster randomized trials in contemporary health systems. *Clinical Trials*. 2015;12(3):276-86. PMC4498459.
- 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.
- Bor J, Moscoe E, Mutevedzi P, Newell ML, Barnighausen T. Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology*. 2014;25(5):729-37. PMC4162343.

References

- Secondary References (cont.)
 - Bor J, Moscoe E, Barnighausen T. Three approaches to causal inference in regression discontinuity designs. *Epidemiology*. 2015;26(2):e28-30.
 - Campbell MJ, Walters SJ. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research*. Chichester: John Wiley & Sons Ltd.; 2014.
 - Cornfield J. Randomization by group: a formal analysis. *American Journal of Epidemiology*. 1978;108(2):100-2.
 - Cook TD, Campbell DT. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Chicago, IL: Rand McNally College Publishing Company; 1979.
 - Crespi CM, Maxwell AE, Wu S. Cluster randomized trials of cancer screening interventions: are appropriate statistical methods being used? *Contemporary Clinical Trials*. 2011;32(4):477-84. PMC3104062.
 - Debar LL, Kindler L, Keefe FJ, Green CA, Smith DH, Deyo RA, Ames K, Feldstein A. A primary care-based interdisciplinary team approach to the treatment of chronic pain utilizing a pragmatic clinical trials framework. *Transl Behav Med*. 2012;2(4):523-30. PMC3578318.

References

- Secondary References (cont.)
 - Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. London: Arnold; 2000.
 - Donner A, Taljaard M, Klar N. The merits of breaking the matches: a cautionary tale. *Statistics in Medicine*. 2007;26(9):2036-51.
 - Flay BR. Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine*. 1986;15(5):451-74.
 - 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.
 - Hade EM, Murray DM, Pennell ML, Rhoda D, Paskett ED, Champion VL, Crabtree BF, Dietrich A, Dignan MB, Farmer M, Fenton JJ, Flocke S, Hiatt RA, Hudson SV, Mitchell M, Monahan P, Shariff-Marco S, Slone SL, Stange K, Stewart SL, Strickland PA. Intraclass correlation estimates for cancer screening outcomes: estimates and applications in the design of group-randomized cancer screening studies. *Journal of the National Cancer Institute Monographs*. 2010;2010(40):97-103. PMC2924625.

References

■ Secondary References (cont.)

- Hayes RJ, Moulton LH. Cluster Randomised Trials. Boca Raton, FL: Taylor & Francis Group, LLC; 2009.
- Hughes JP, Granston TS, Heagerty PJ. Current issues in the design and analysis of stepped wedge trials. *Contemporary Clinical Trials*. 2015;45(Pt A):55-60. PMC4639463.
- Jarvik JG, Comstock BA, James KT, Avins AL, Bresnahan BW, Deyo RA, Luetmer PH, Friedly JL, Meier EN, Cherkin DC, Gold LS, Rundell SD, Halabi SS, Kallmes DF, Tan KW, Turner JA, Kessler LG, Lavalley DC, Stephens KA, Heagerty PJ. Lumbar Imaging With Reporting Of Epidemiology (LIRE)-Protocol for a pragmatic cluster randomized trial. *Contemporary Clinical Trials*. 2015;45(Pt B):157-63. PMC4674321.
- 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.

References

■ Secondary References (cont.)

- Johnson KE, Tachibana C, Coronado GD, Dember LM, Glasgow RE, Huang SS, Martin PJ, Richards J, Rosenthal G, Septimus E, Simon GE, Solberg L, Suls J, Thompson E, Larson EB. A guide to research partnerships for pragmatic clinical trials. *BMJ*. 2014;349:g6826.
- 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. doi: 10.1002/sim.6813.
- Moscoe E, Bor J, Barnighausen T. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. *Journal of Clinical Epidemiology*. 2015;68(2): 122-33.
- Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clinical Trials*. 2004;1(3):297-305.
- Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review*. 2003;27(1): 79-103.

References

- Secondary References (cont.)
 - Murray DM, Catellier DJ, Hannan PJ, Treuth MJ, Stevens J, Schmitz K, Rice J, Conway T. School-level intraclass correlation for physical activity in adolescent girls. *Medicine and Science in Sports and Exercise*. 2004;36(5):876-82. PMC2040294.
 - 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.
 - 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.
 - Murray DM, Hannan PJ, Wolfinger RD, Baker WL, Dwyer JH. Analysis of data from group-randomized trials with repeat observations on the same groups. *Statistics in Medicine*. 1998;17(14):1581-600.
 - Murray DM, Pals SP, Blitstein JL, Alfano CM, Lehman J. Design and analysis of group-randomized trials in cancer: a review of current practices. *J Natl Cancer Inst*. 2008;100(7):483-91.

References

■ Secondary References (cont.)

- Murray DM, Pennell M, Rhoda D, Hade E, Paskett ED. Designing studies that would address the multilayered nature of health care. *Journal of the National Cancer Institute Monographs*, 2010, 40:90-96.
- Murray DM, Rooney BL, Hannan PJ, Peterson AV, Ary DV, Biglan A, Botvin GJ, Evans RI, Flay BR, Futterman R, Getz JG, Marek PM, Orlandi M, Pentz MA, Perry CL, Schinke SP. Intraclass correlation among common measures of adolescent smoking: estimates, correlates, and applications in smoking prevention studies. *American Journal of Epidemiology*. 1994;140(11):1038-50.
- Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health*. 2004;94(3):423-32.
- O'Keeffe AG, Geneletti S, Baio G, Sharples LD, Nazareth I, Petersen I. Regression discontinuity designs: an approach to the evaluation of treatment efficacy in primary care using observational data. *BMJ*. 2014;349:g5293.

References

■ Secondary References (cont.)

- Pals SP, Murray DM, Alfano CM, Shadish WR, Hannan PJ, MStat, 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.
- Pennell ML, Hade EM, Murray DM, Rhoda DA. Cutoff designs for community-based intervention studies. *Stat Med*. 2011;30(15):1865-82.
- Pratt CA, Boyington J, Esposito L, Pemberton VL, Bonds D, Kelley M, Yang S, Murray D, Stevens J. Childhood Obesity Prevention and Treatment Research (COPTR): interventions addressing multiple influences in childhood and adolescent obesity. *Contemporary Clinical Trials*. 2013;36(2):406-13.
- Rhoda DA, Murray DM, Andridge RR, Pennell ML, Hade EM. Studies with staggered starts: multiple baseline designs and group-randomized trials. *Am J Public Health*. 2011;101(11):2164-9.

References

- Secondary References (cont.)
 - Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, Bauck AE, Cifelli D, Smerek MM, Dickerson J, Laws RL, Madigan RA, Rusincovitch SA, Kluchar C, Califf RM. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *Journal of the American Medical Informatics Association*. 2013;20(e2):e226-31. PMC3861929.
 - Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *Journal of Chronic Diseases*. 1967;20(8):637-48.
 - 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.
 - Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Company; 2002.
 - Swallow WH, Monahan JF. Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics*. 1984;26(1):47-57.

References

- Secondary references (cont.)
 - 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.
 - Zucker DM. An analysis of variance pitfall: The fixed effects analysis in a nested design. *Educational and Psychological Measurement*. 1990;50(4):731-8.