Design and Analysis of Group-Randomized Trials

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

2015 OBSSR Summer Institute
July 20, 2015

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).
  - Efficacy trials evaluate an intervention under carefully controlled conditions.
  - Effectiveness trials evaluate an intervention under real-world conditions.

- Schwartz, D., & Lellouch, J. Explanatory and pragmatic attitudes in therapeutical trials. Journal of Chronic Diseases, 1967, 20(8), 637-648.
- Cochrane, A.L. Effectiveness and efficacy: random reflections on health services. Nuffield Provincial Hospitals Trust, London, 1971. (cited in Flay, Brian R. Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. Preventive Medicine, 1986, 15(5), 451-474.)

Methodological Considerations

- Pragmatic trials do not necessarily require a different set of research designs, measures, analytic methods, etc.
- As always, the choice of methods depends on the research question.
- The research question dictates
  - the intervention, target population, and variables of interest,
  - which dictate the setting, research design, measures, and analytic methods.
- Randomized trials will provide the strongest evidence.
  - Which kind of randomized trial will depend on the research question.
- Alternatives to randomized trials are also available.

Three Kinds of Randomized Trials

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

Examples

- Group-randomized trials
  - Health Care Systems Collaboratory
    - 7 pragmatic trials conducted in collaboration with health care systems
    - 5 are group-randomized trials
      - Hospital acquired infections, CRC screening, mortality in dialysis patients, healthcare utilization in spinal injuries, chronic pain management
- Individually randomized group treatment trials
  - Childhood Obesity Prevention and Treatment Research (COPTR)
    - Two prevention studies targeting young children
    - Two treatment studies targeting youth
    - Four separate studies, with separate designs, interventions, evaluations
    - All involve substantial participant interaction post-randomization

Impact on the Design

- Randomized clinical trials
  - There is usually good opportunity for randomization to distribute all potential sources of bias evenly.
  - If well executed, bias is not usually a concern.
- Individually randomized group treatment trials
  - Non-random assignment to small groups may create bias.
  - Bias can be more of a concern in IRGTs than in RCTs.
- Group-randomized trials
  - GRTs often involve a limited number of groups.
  - In any single realization, there is limited opportunity for randomization to distribute all potential sources of bias evenly.
  - Bias is more of a concern in GRTs than in RCTs.

Impact on the Analysis

- 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.
- That connection will create a positive intraclass correlation that reflects an extra component of variance attributable to the group.

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

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

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

Impact on the Analysis

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

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

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

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

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

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

Impact on the Analysis in a GRT

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

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

  - When group membership is not established by random assignment,

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

  - Or equivalently,

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

Impact on the Analysis

- 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 an issue in a GRT.
  - This can be an issue in an IRGT, especially if there are small groups in all study conditions.
- 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 limit power, so they must be considered at the design stage.

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, J. (1978). Randomization by group: a formal analysis. *American Journal of Epidemiology*, 108(2), 100-102.

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

Impact on the Analysis

- We can estimate the effect of the ICC as:

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

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

- Scott, A. J., & Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77(380), 848-854.

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.
- For studies in which the groups are crossed with the levels of the exposure of interest (most observational studies)…
  - $ICC_x \approx ICC_y$.
- For studies in which 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:

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

The Need for GRTs and IRGTs

- 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 without contamination
- An IRGT is the best comparative design whenever...
  - Individual randomization is possible without contamination
  - There are good reasons to deliver the intervention in small 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.

Potential Threats to Internal Validity

- Four primary threats:
  - Selection
  - History and differential history
  - Maturation and differential maturation
  - Contamination

Strategies to Limit Threats to Internal Validity

- Randomization
- A priori matching or stratification
  - Of groups in GRTs, of members in IRGTs
- 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

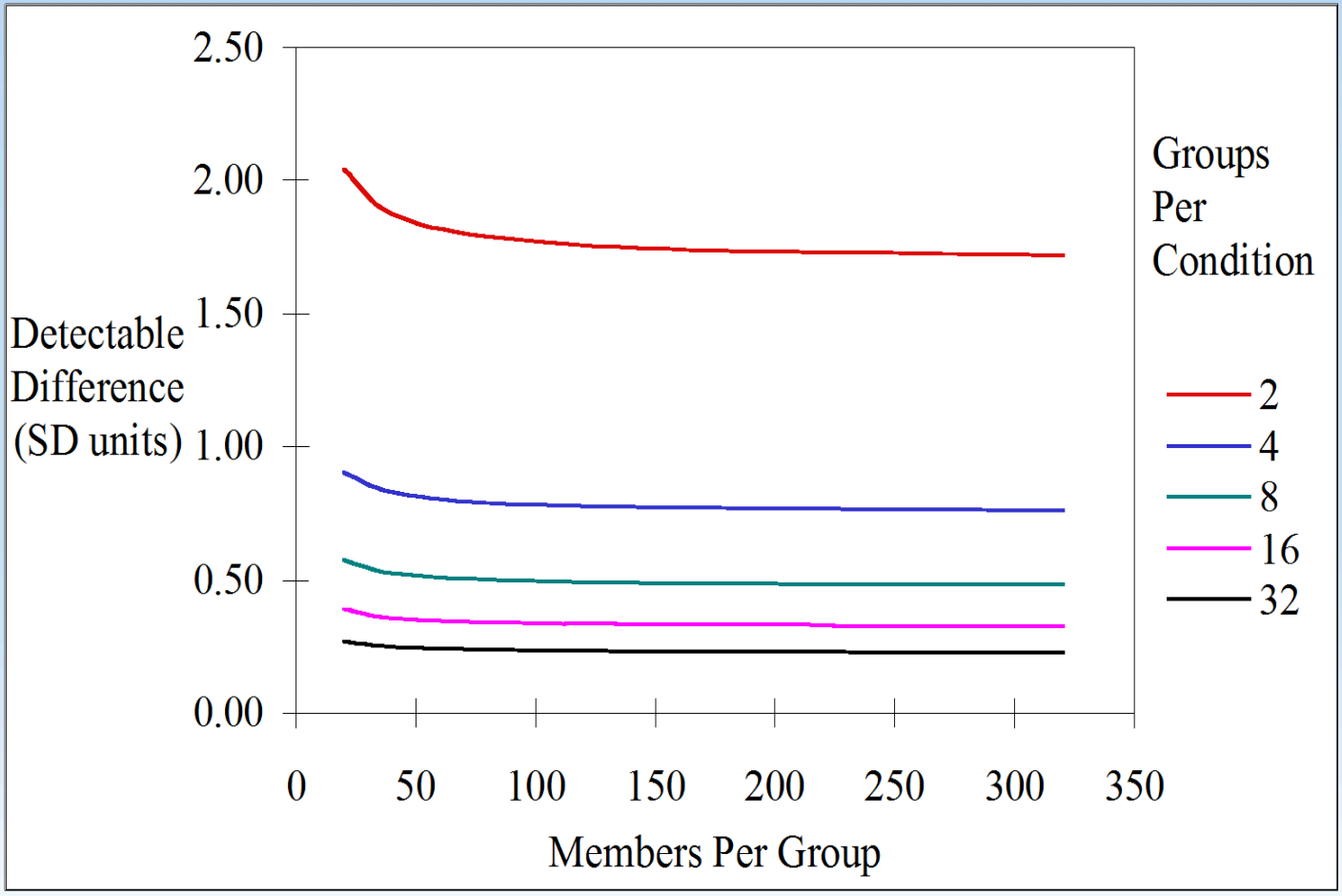
- The variance of the condition mean in a GRT is:

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

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

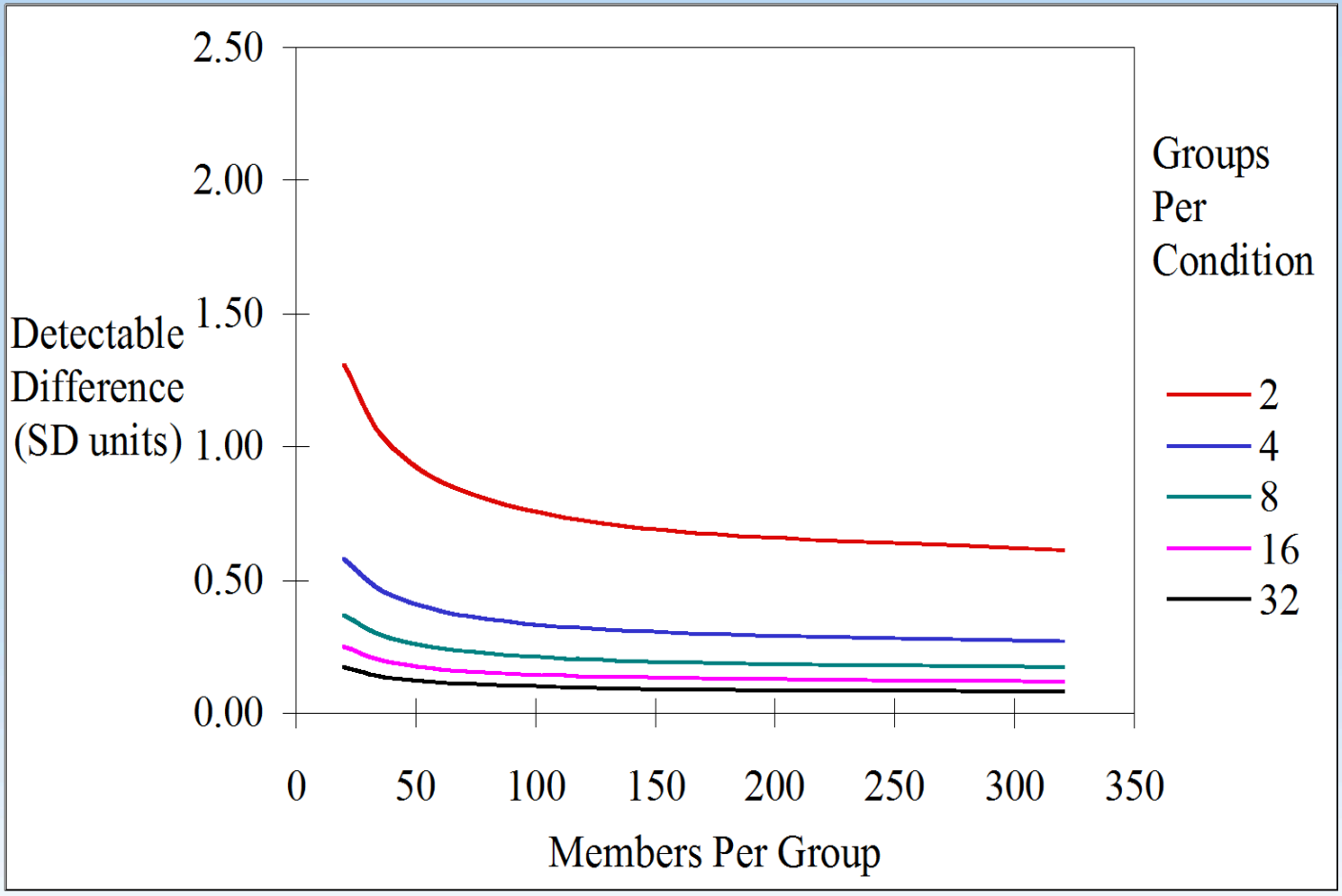
Strategies to Improve Precision

- Increased replication (ICC=0.100)



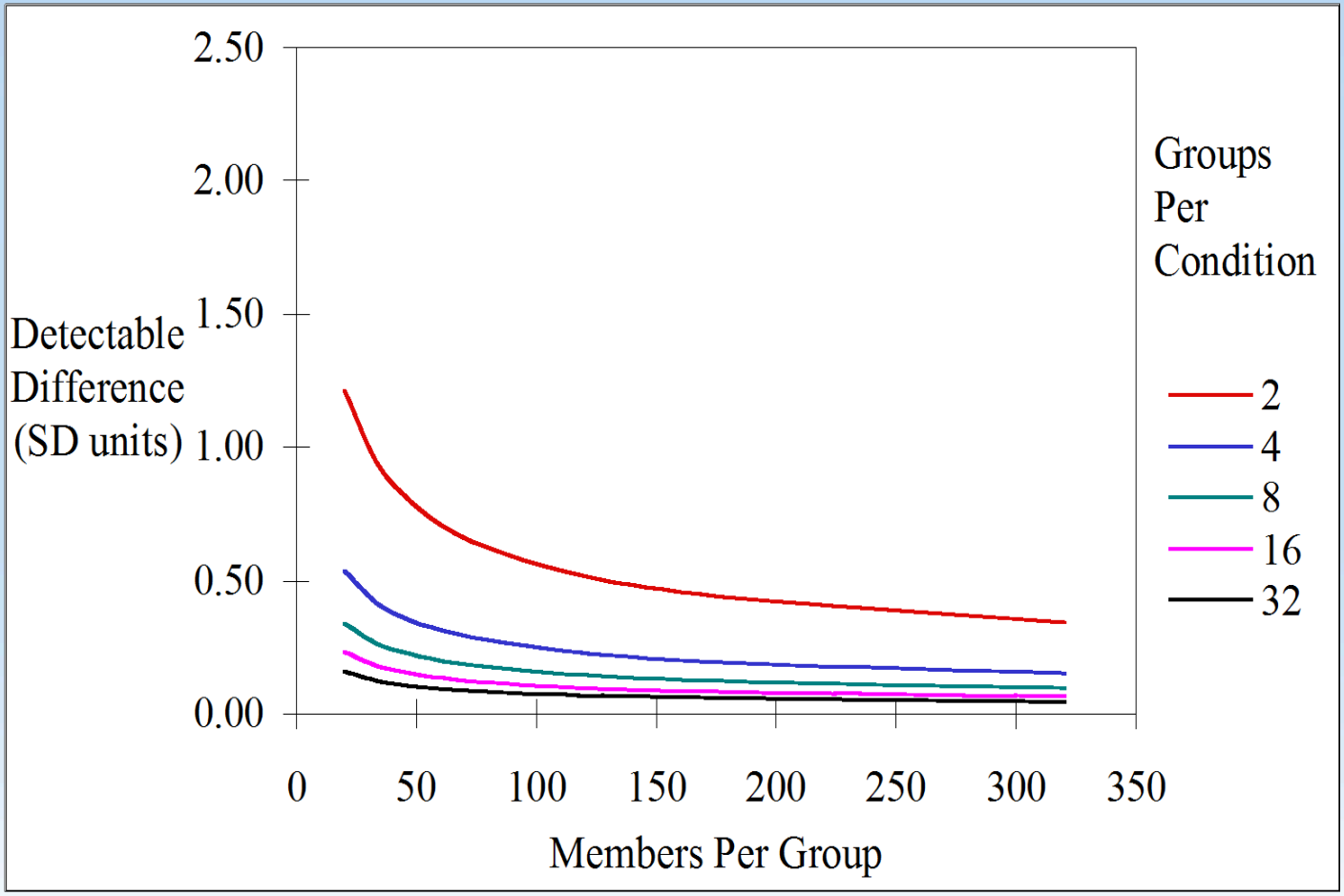
Strategies to Improve Precision

- Reduced ICC (ICC=0.010)



Strategies to Improve Precision

- The law of diminishing returns (ICC=0.001)



Power for Group-Randomized Trials

- The usual methods must be adapted to reflect the nested design
  - The variance is greater in a GRT due to the expected ICC.
  - df should be based on the number of groups, not the number of members.
- A good source on power is Chapter 9 in Murray (1998).
- Many papers now report ICCs and show how to plan a GRT.
  - cf. Murray & Blitstein, 2003 and Murray et al., 2004.
- Power in GRTs is tricky, and investigators are advised to get help from someone familiar with these methods.
- Power for IRGTs is often even trickier, and the literature is more limited.
  - cf. Pals et al. 2008.

Power - RCTs vs GRTs

- A simple RCT

$$n = \frac{2\sigma_y^2\left(t_{\alpha/2} + t_{\beta}\right)^2}{\Delta^2}, \quad df = 2(n-1)$$

- A simple GRT

$$g = \frac{2\sigma_y^2\left(t_{\alpha/2} + t_{\beta}\right)^2\left(1 + \left(m-1\right)ICC_{m:g:c}\right)}{m\Delta^2}, \quad df = 2(g-1)$$

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 Having 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 that these methods have the nominal Type I error rate across a wide range of conditions common in GRTs.

Preferred Analytic Strategies for Designs Having More Than Two Time Intervals

- Random coefficients models
  - Sometimes 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 has an inflated Type I error rate if those trends are heterogeneous.
  - Random coefficients models allow for heterogeneity of those trends.
    - Random coefficients models have the nominal Type I error rate across a wide range of conditions common in GRTs.
  - Random coefficients models are used increasingly in the evaluation of public health interventions.
    - Examples include NCI's Project ASSIST and NHLBI's REACT.

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.
  - Randomization does ensure the nominal Type I error rate, even when very few heterogeneous groups are assigned to each condition.
  - Programs for randomization tests are available in print and on the web.
- 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, where model-based methods provide parameter estimates, standard errors, etc.
  - Regression adjustment for covariates requires many of the same assumptions as the model-based tests.

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

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

What About Fixed-Effect Methods in Two Stages?

- Introduced as the first solution to the unit of analysis problem 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.

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.

- 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 causal inference is important.
  - It may still be helpful for preliminary studies.

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.
  - 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.
- Analyses that ignore the ICC risk an inflated Type I error rate.
  - 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.

Closing Thoughts

- Pragmatic trials are increasingly of interest and many involve group or cluster randomization.
- 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 methodologists.

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.

- 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.

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

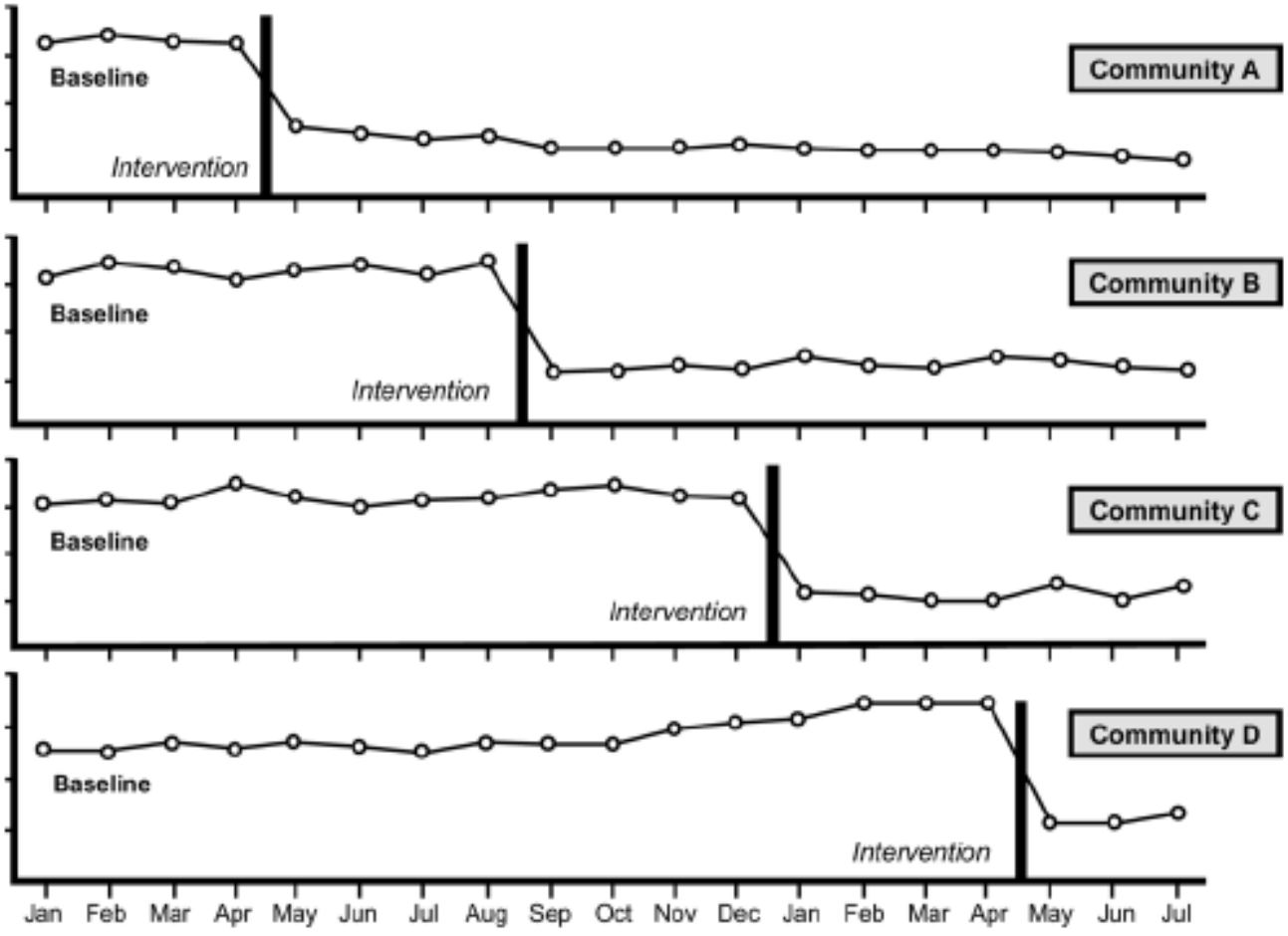


Figure 1. Hypothetical example of a multiple baseline design used to assess behavior change following an intervention in four communities.

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.

- 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.

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.

Time Series Designs

- Goldberg et al. evaluated a cancer screening program.
  - One clinic served as the intervention site, a second as the control site.
  - Screening rates were monitored weekly before and after the intervention.
  - The analysis was based on least-square regression after preliminary analyses revealed no evidence of autocorrelation.
  - Causal inference is not possible, but the results can set the stage for an efficacy study.
    - The analysis could not separate variation due to clinic from variation due to community, and so could not provide a valid test for the intervention.

- Goldberg HI et al. A controlled time-series trial of clinical reminders: using computerized firm systems to make quality improvement research a routine part of mainstream practice. Health Serv Res. 2000;37(7):1519-34.

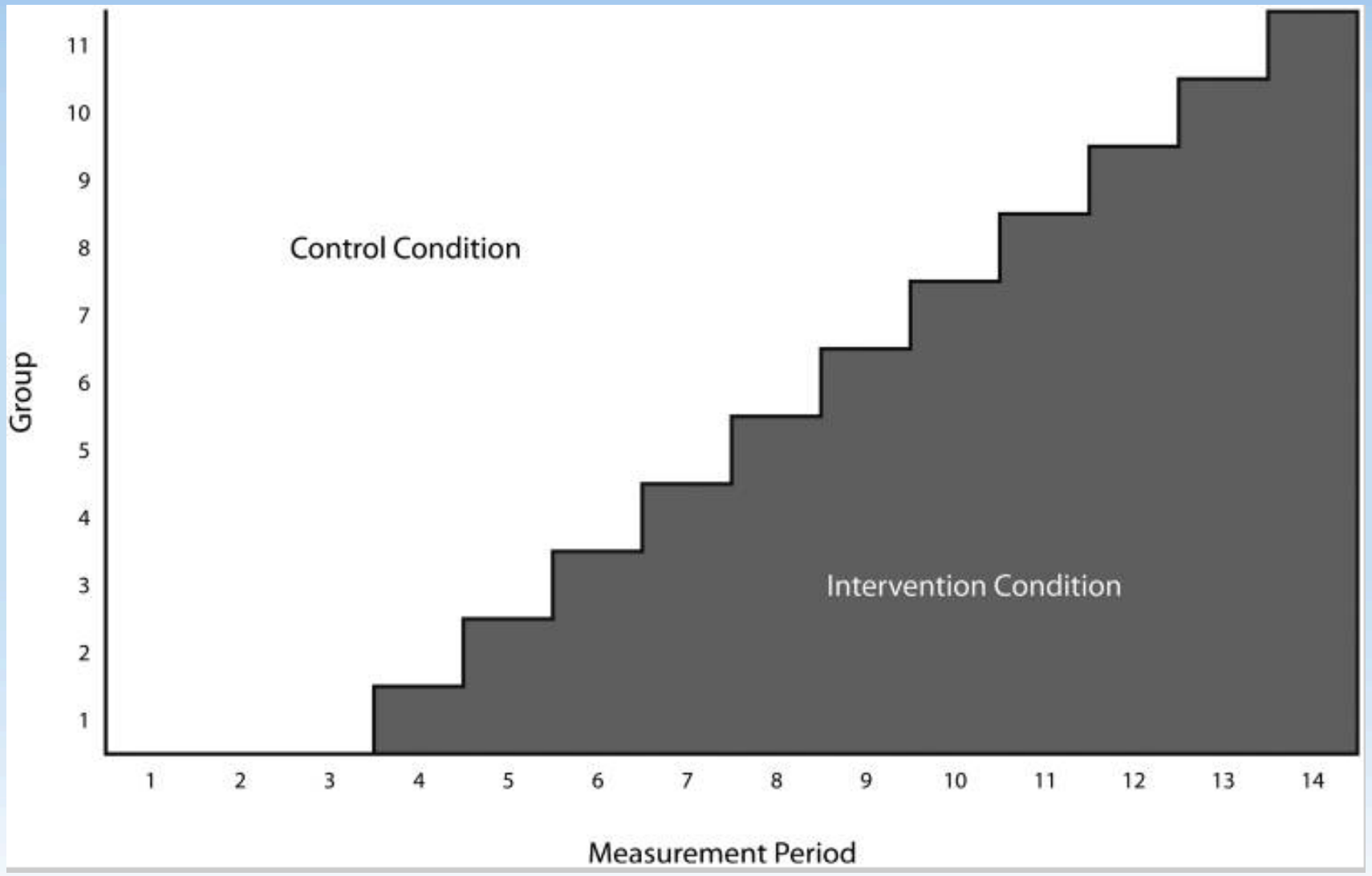
Quasi-Experimental Designs

- 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.

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.
- Example
  - Pragmatic Trial of Lumbar Image Reporting with Epidemiology (LIRE), Jeffery Jarvik PI, HCS Collaboratory Project

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.

- 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.

Regression Discontinuity Designs

- Groups or individuals are assigned to conditions based on a score, often reflecting the need for the intervention.
- The analysis models the relationship between the assignment variable and the outcome.
  - The difference in intercepts at the cutoff is the intervention effect.

Regression Discontinuity Design

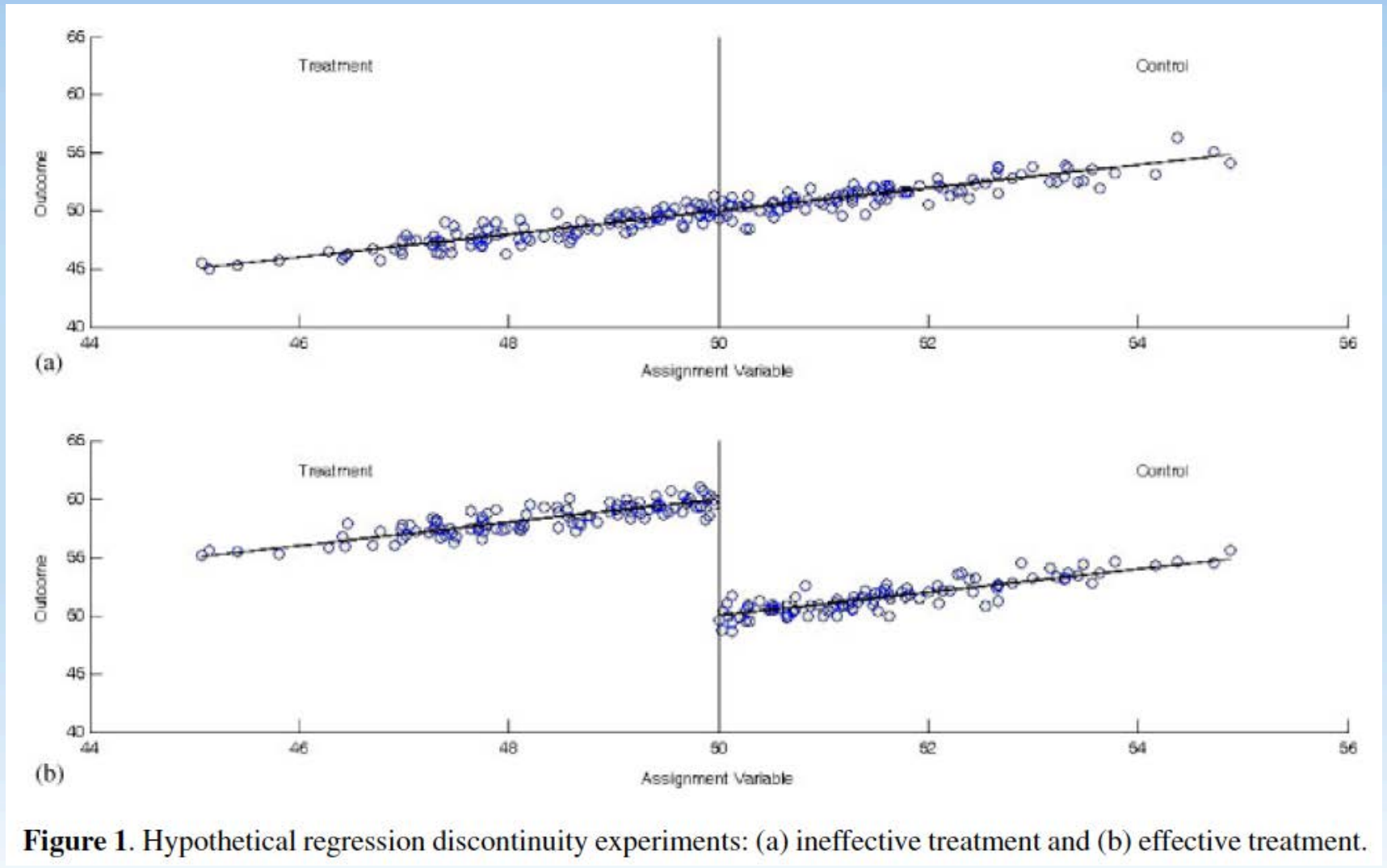


Figure 1. Hypothetical regression discontinuity experiments: (a) ineffective treatment and (b) effective treatment.

Regression Discontinuity Design

- Because assignment is fully explained by the assignment variable, proper modeling supports causal inference.
  - Rubin, Assignment to Treatment Group on the Basis of a Covariate, Journal of Educational and Behavioral Statistics, 1977, 2:1-26.
- 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.

- Pennell ML, Hade EM, Murray DM, Rhoda DA. Cutoff designs for community-based intervention studies. Statistics in Medicine 2011;30(15):1865-1882.

References

- Primary References
  - Murray, D.M. Design and Analysis of Group-Randomized Trials. New York: Oxford University Press, 1998.
  - 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):90-6.
- Secondary References
  - 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.
  - Pennell ML, Hade EM, Murray DM, Rhoda DA. Cutoff designs for community-based intervention studies. Stat Med. 2011;30(15):1865-82.

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.
  - 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.
  - 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.