Design and Analysis of Studies to Evaluate Multilevel Interventions in Public Health and Medicine

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

Medicine: Mind the Gap Seminar
October 23, 2015

National Institutes of Health
Office of Disease Prevention

Multilevel Interventions

- Multilevel interventions address more than one level of influence for the targeted outcome.
- Multilevel interventions pose special challenges in terms of design and analysis.
  - Respondents who share the same source for any level of influence will share some physical, social, or other connection.
  - Such connections create a positive intraclass correlation among the observations taken from those respondents.
  - That correlation invalidates the usual analytic procedures.
  - This must be considered in the planning stage to ensure a valid analysis and adequate power.
- Many different design and analytic alternatives have been proposed for the evaluation of multilevel interventions.

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, etc.
    - Also known as cluster-randomized trials

Impact on the Design

- Randomized clinical trials and individually randomized group-treatment 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.
- 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

- 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 (ICC) that reflects extra variation attributable to the group:

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

Impact on the Analysis

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

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

  - When group membership is not established by random assignment,

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

  - Or equivalently,

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

Impact on the Analysis

- 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 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, or
  - cannot be delivered to individuals without contamination.
- An IRGT is the best comparative design whenever...
  - individual randomization is possible without contamination, but
  - there are good reasons to deliver the intervention in small groups.

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

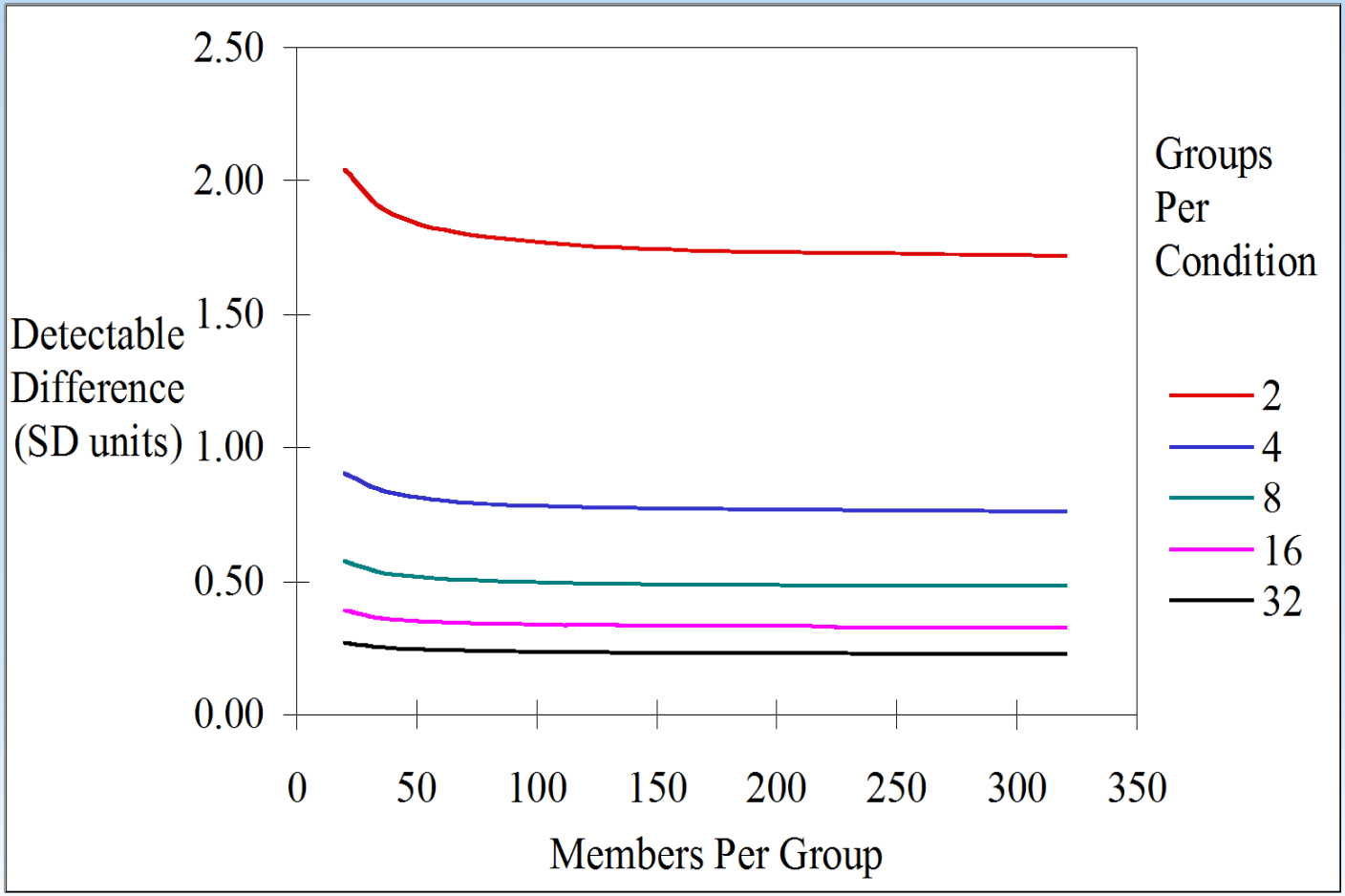
- The variance of the condition mean in a GRT is:

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

- 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

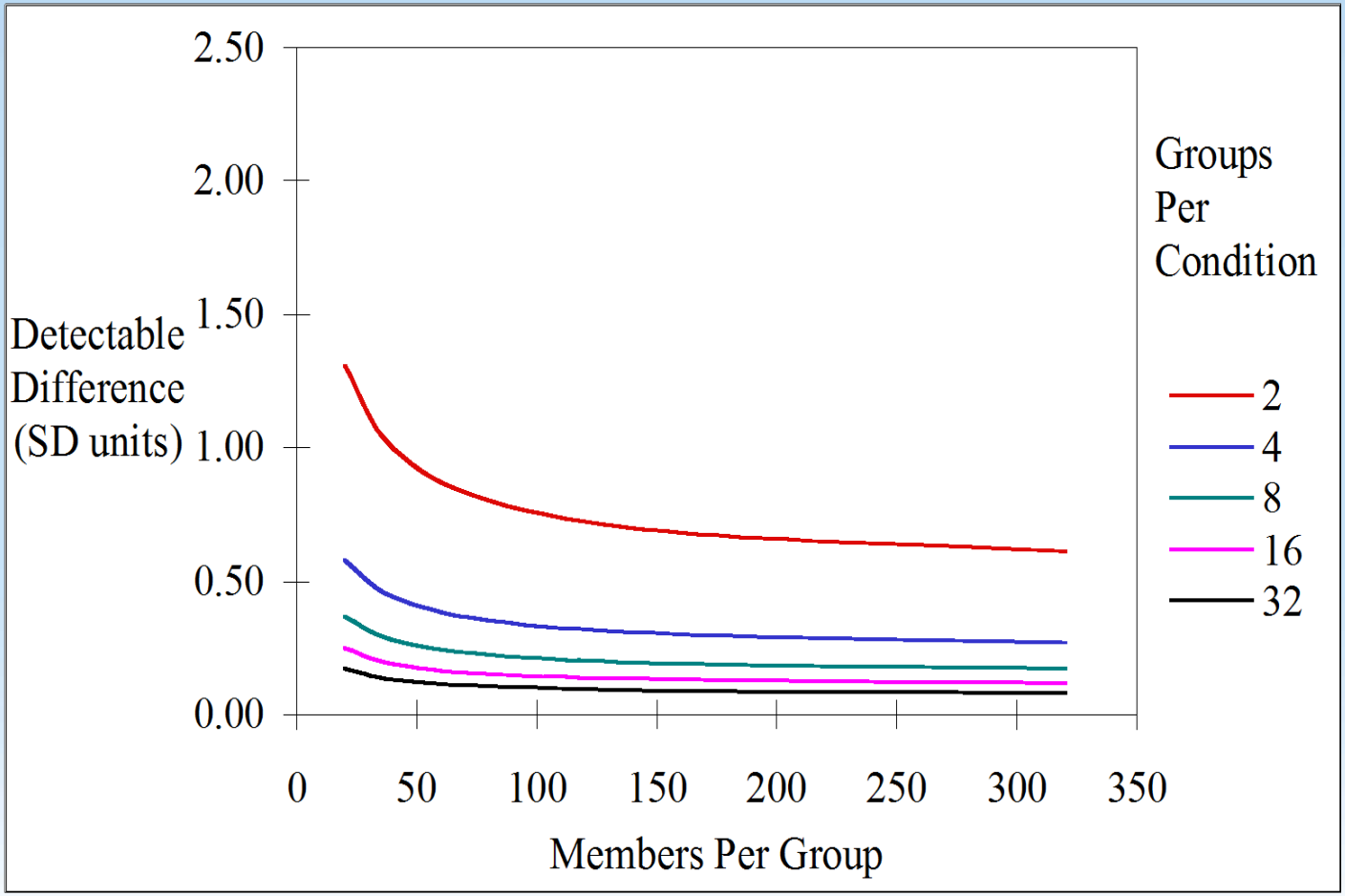
Strategies to Improve Precision

- Increased replication (ICC=0.100)



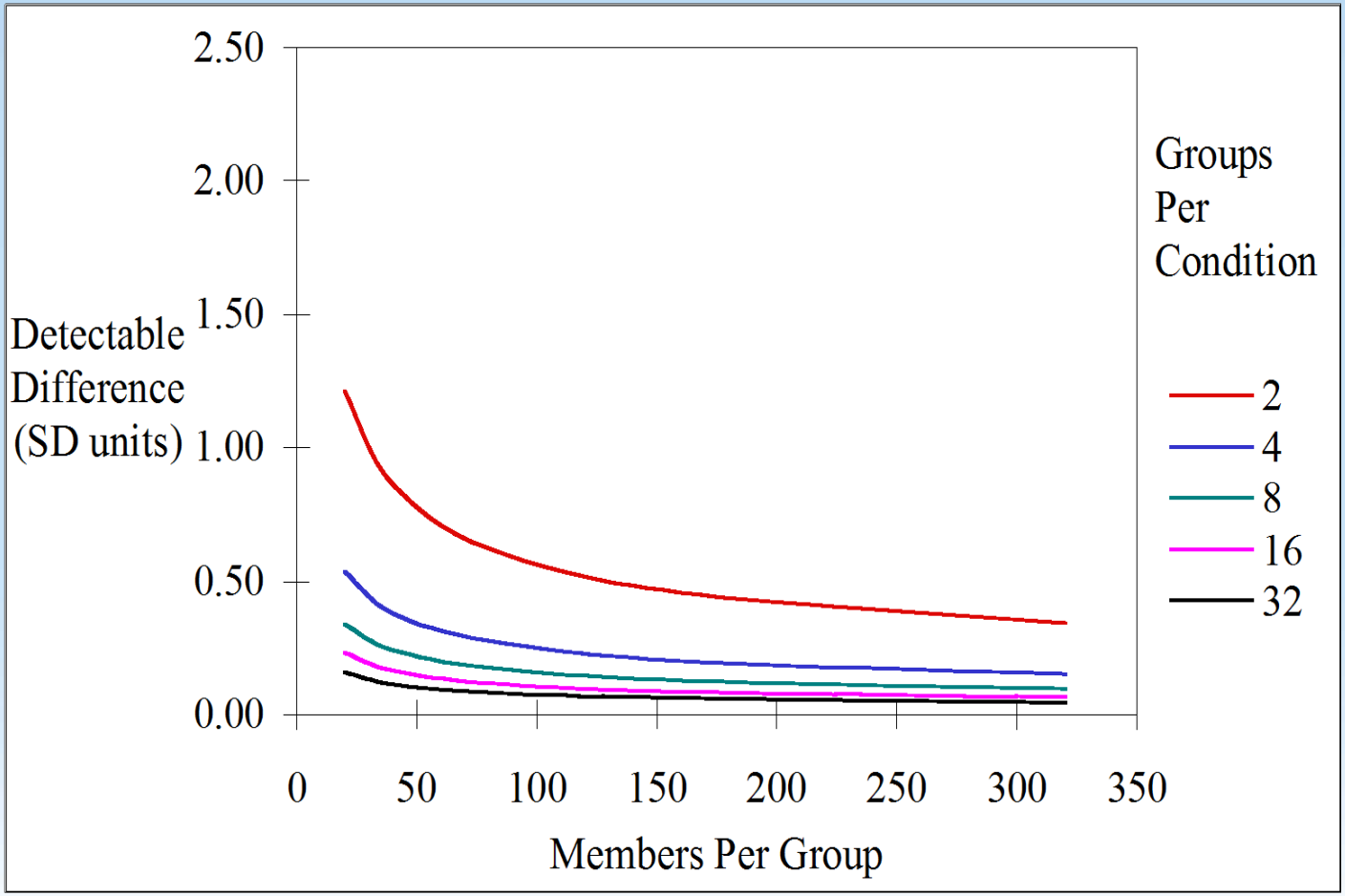
Strategies to Improve Precision

- Reduced ICC (ICC=0.010)



Strategies to Improve Precision

- The law of diminishing returns (ICC=0.001)



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.

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.
- 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.

What About Alternative Designs?

- Many alternatives to GRTs have been proposed.
  - Multiple baseline designs
  - Time series designs
  - Quasi-experimental (QE) designs
  - Dynamic wait-list or stepped-wedge designs
  - Regression discontinuity (RD) 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

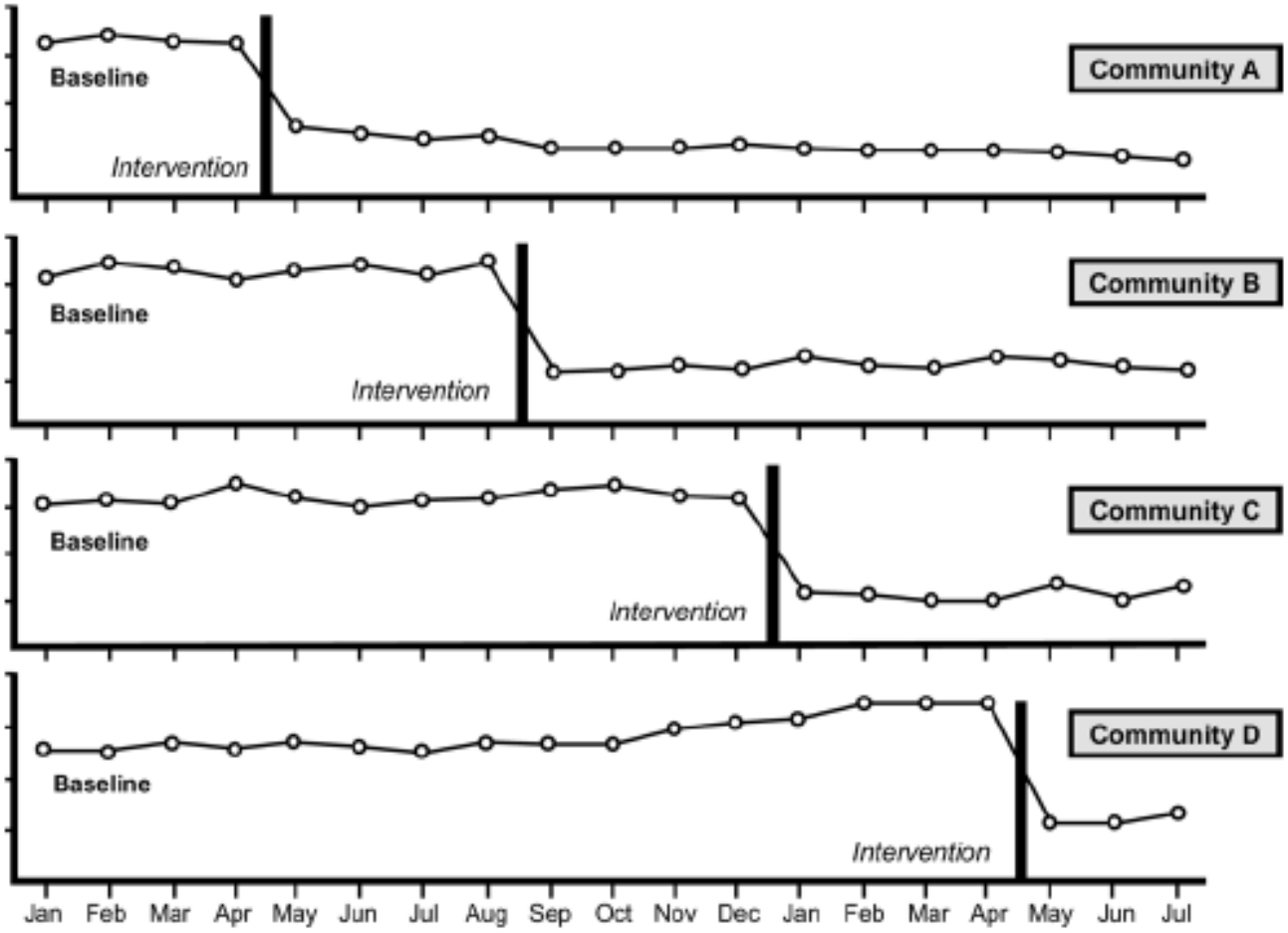


Figure 1. Hypothetical example of a multiple baseline design used to assess behavior change following an intervention in four communities.

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 with an archival data collection system.
  - Could be a strong approach with archival data on many groups.
- May require several cycles of data.

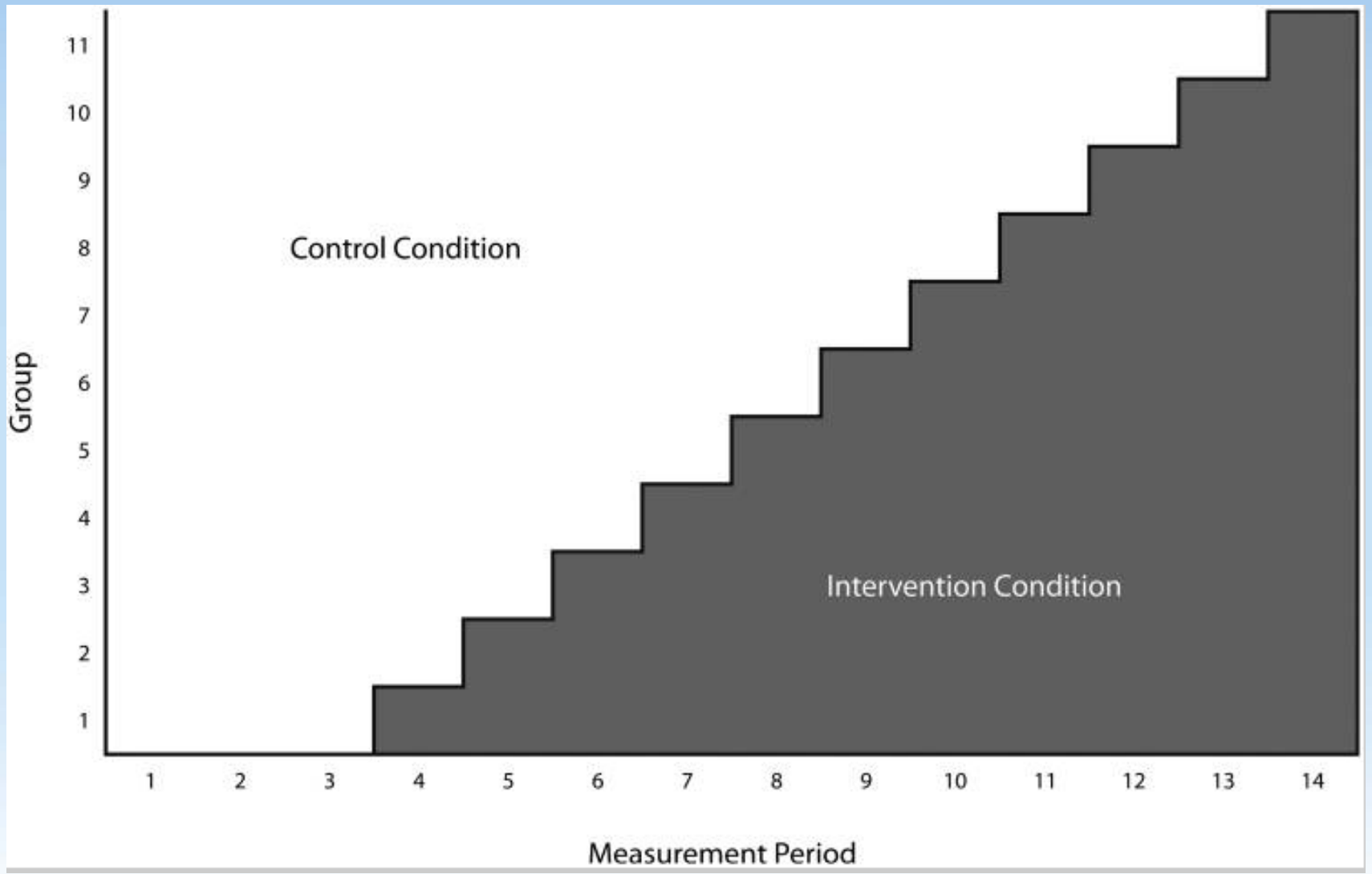
Quasi-Experimental Designs

- 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.

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.

Stepped Wedge Design



Stepped Wedge Design

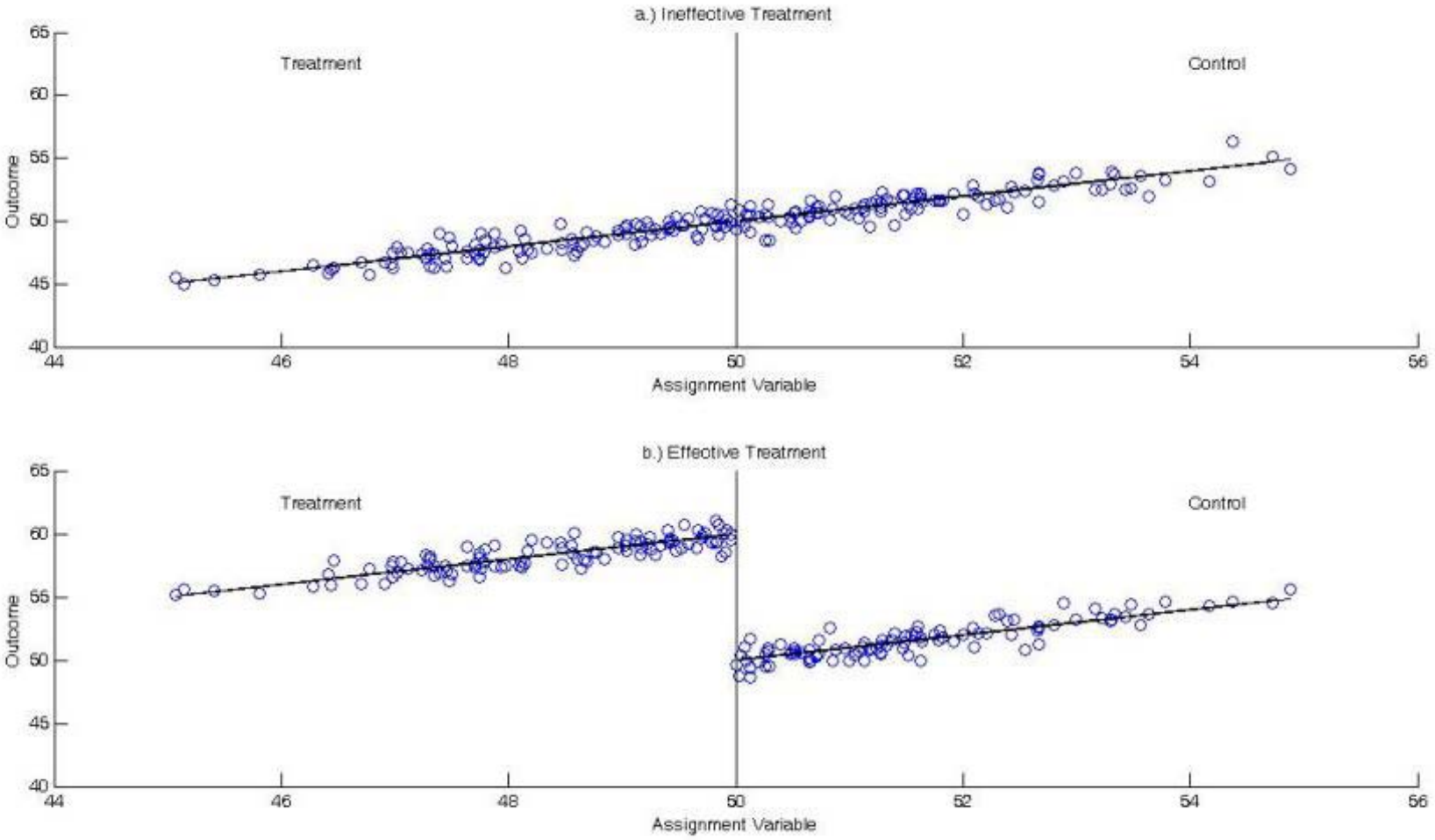
- The analysis estimates a weighted average intervention effect across the intervals.
  - Assumes that the intervention effect is rapid and lasting.
  - Not very sensitive to intervention effects that develop gradually or fade over time.
- These designs can be more efficient, but usually take longer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



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 an 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.

Closing Thoughts

- GRTs, IRGTs, stepped wedge, and regression discontinuity designs can provide the strongest evidence for causal inference if implemented and analyzed carefully.
  - Consider extra variation and limited df at the design stage.
  - Randomize with stratification on the baseline value of the primary outcome and group size.
    - Consider regression discontinuity if randomization is not possible.
  - Blind evaluation staff, to the extent possible.
  - Analyze to account for extra variation, limited df, and member-level imbalance.
- Other approaches can also provide evidence for causal inference, but rely on logic as much as statistics and face more threats to causal inference.

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. *JNCI Monographs*. 2010(40):90-6.
- Secondary References
  - 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. *Stat Med*. 2015. doi: 10.1002/sim.6565.
  - 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. *Stat Med*. 2014. doi: 10.1002/sim.6083

References

- Secondary References (cont.)
  - Roberts C, Walwyn R. Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. Statistics in Medicine. 2013. doi: 10.1002/sim.5521.
  - 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.
  - 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.

Medicine: Mind the Gap Seminar

Design and Analysis of Studies to Evaluate Multilevel Interventions in Public Health and Medicine

David M. Murray, Ph.D.

Associate Director for Prevention

Director, Office of Disease Prevention

prevention@mail.nih.gov

https://prevention.nih.gov/

National Institutes of Health

Office of Disease Prevention