Pragmatic and Group-Randomized Trials in Public Health and Medicine
Part 4: Power and Sample Size

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

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

Target Audience

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

Learning Objectives

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

Organization of the Course

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

Power for Group-Randomized Trials

- The usual methods must be adapted for 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; Moerbeek & Teerenstra, 2016.
  - Recent review articles include Gao et al. (2015) and Rutterford et al. (2015).

- Murray DM. Design and Analysis of Group-Randomized Trials. New York, NY: Oxford University Press; 1998.
- Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. London: Arnold; 2000.
- Hayes RJ, Moulton LH. Cluster Randomised Trials. Boca Raton, FL: CRC Press; 2009.
- 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.
- Moerbeek M, Teerenstra S. Power analysis of trials with multilevel data. Boca Raton: CRC Press; 2016.
- Gao F, Earnest A, Matchar DB, Campbell MJ, Machin D. Sample size calculations for the design of cluster randomized trials: A summary of methodology. Contemporary Clinical Trials. 2015;42:41-50.
- Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. International Journal of Epidemiology. 2015;44(3):1051-67. PMC4521133.

Power for Group-Randomized Trials

- 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; Heo et al., 2014; Moerbeek & Teerenstra, 2016).

- Pals SP, Murray DM, Alfano CM, Shadish WR, Hannan PJ, Baker WL. Individually randomized group treatment trials: a critical appraisal of frequently used design and analytic approaches. American Journal of Public Health. 2008;98(8):1418-24. PMC2446464
- Pals SL, Murray DM, Alfano CM, Shadish WR, Hannan PJ, Baker WL. Erratum. American Journal of Public Health. 2008;98(12):2120.
- Heo M, Litwin AH, Blackstock O, Kim N, Arnsten JH. Sample size determinations for group-based randomized clinical trials with different levels of data hierarchy between experimental and control arms. Statistical Methods in Medical Research. 2014. PMC4329103.
- Moerbeek M, Teerenstra S. Power analysis of trials with multilevel data. Boca Raton: CRC Press; 2016.

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

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

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, 2006)

- Murray DM, Rooney BL, Hannan PJ, et al. 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, Stevens J, Hannan PJ, Catellier DJ, Schmitz KH, Dowda M, Conway TL, Rice JC, Yang S. School-level intraclass correlation for physical activity in sixth grade girls. Medicine and Science in Sports and Exercise. 2006;38(5):926-36. PMC2034369.

Strategies to Reduce Extra Variation

- Effective strategies
  - Regression adjustment for covariates
    - Fixed covariates in non-repeated measures analyses
    - Time-varying covariates in repeated measures analyses
    - This is one of the most effective methods to reduce intraclass correlation and extra variation (Murray & Blitstein, 2003) and will often reduce the ICC by 50-75%.

- Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. Evaluation Review. 2003;27(1):79-103.

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 member.

- 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, 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.

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 have been 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).

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

Estimating ICC

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

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

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:

$$\hat{\Delta} = \sqrt{\hat{\sigma}_{\Delta}^{2}\left(t_{critical:\alpha/2} + t_{critical:\beta}\right)^{2}}$$

$$= \sqrt{2\left(\frac{\sigma_{y}^{2}}{n}\right)\left(t_{critical:\alpha/2} + t_{critical:\beta}\right)^{2}}$$

Detectable Difference

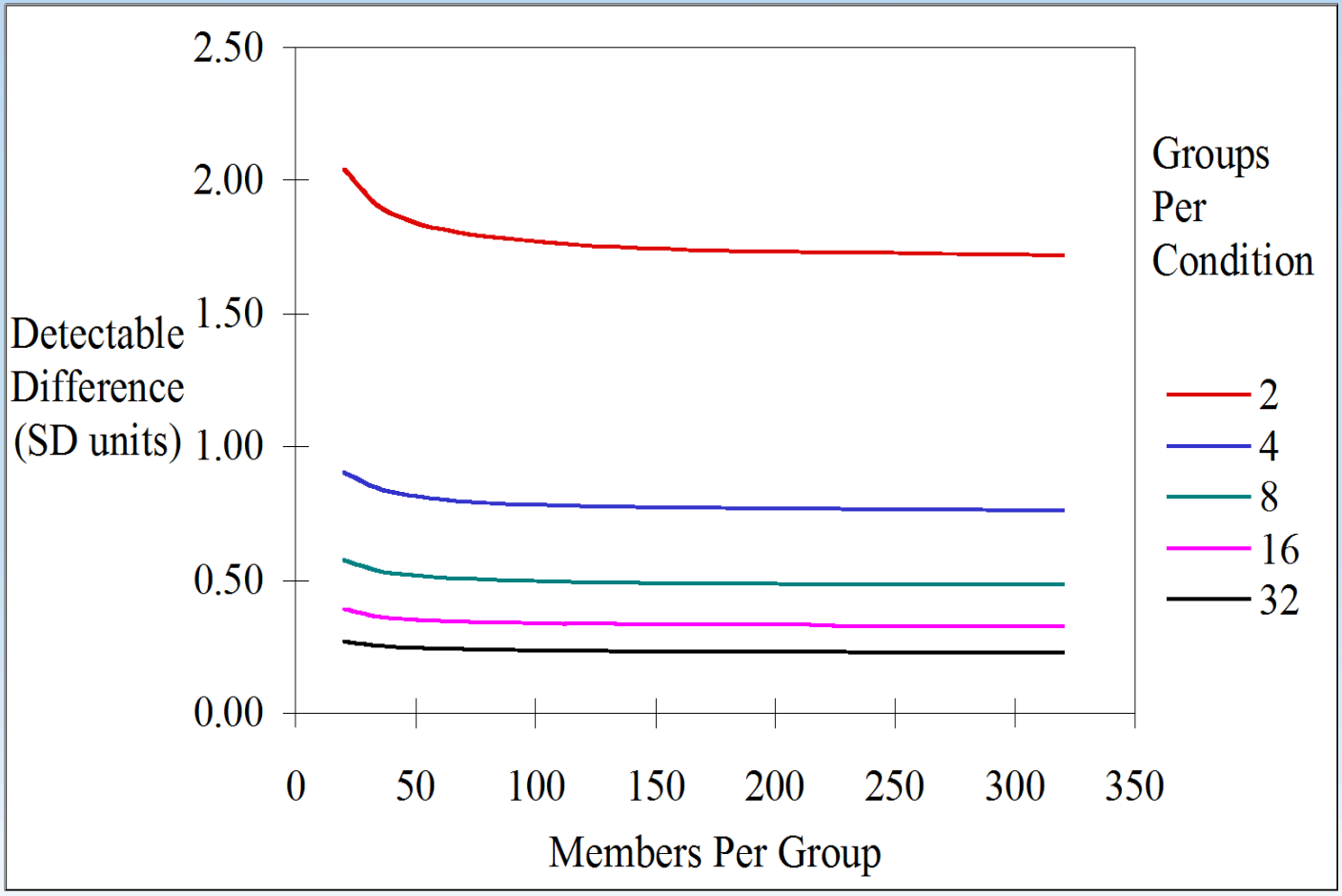
- The seventh step…
  - Calculate sample size, detectable difference, or power based on those estimates.
  - For a one df contrast between two condition means or mean slopes, the detectable difference in a simple GRT is:

$$\hat{\Delta} = \sqrt{\hat{\sigma}_{\Delta}^{2}\left(t_{critical:\alpha/2} + t_{critical:\beta}\right)^{2}}$$

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

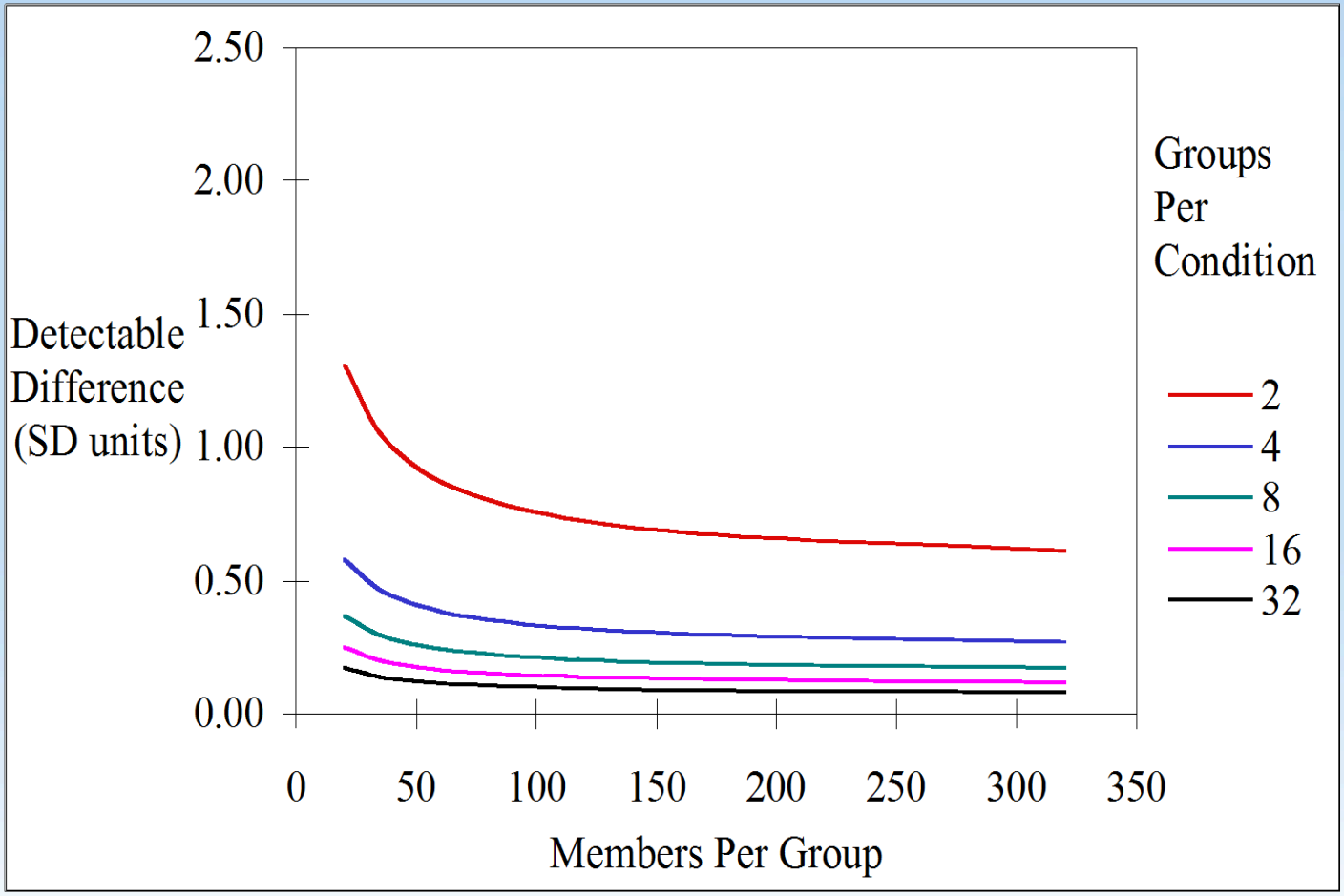
Detectable Difference

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



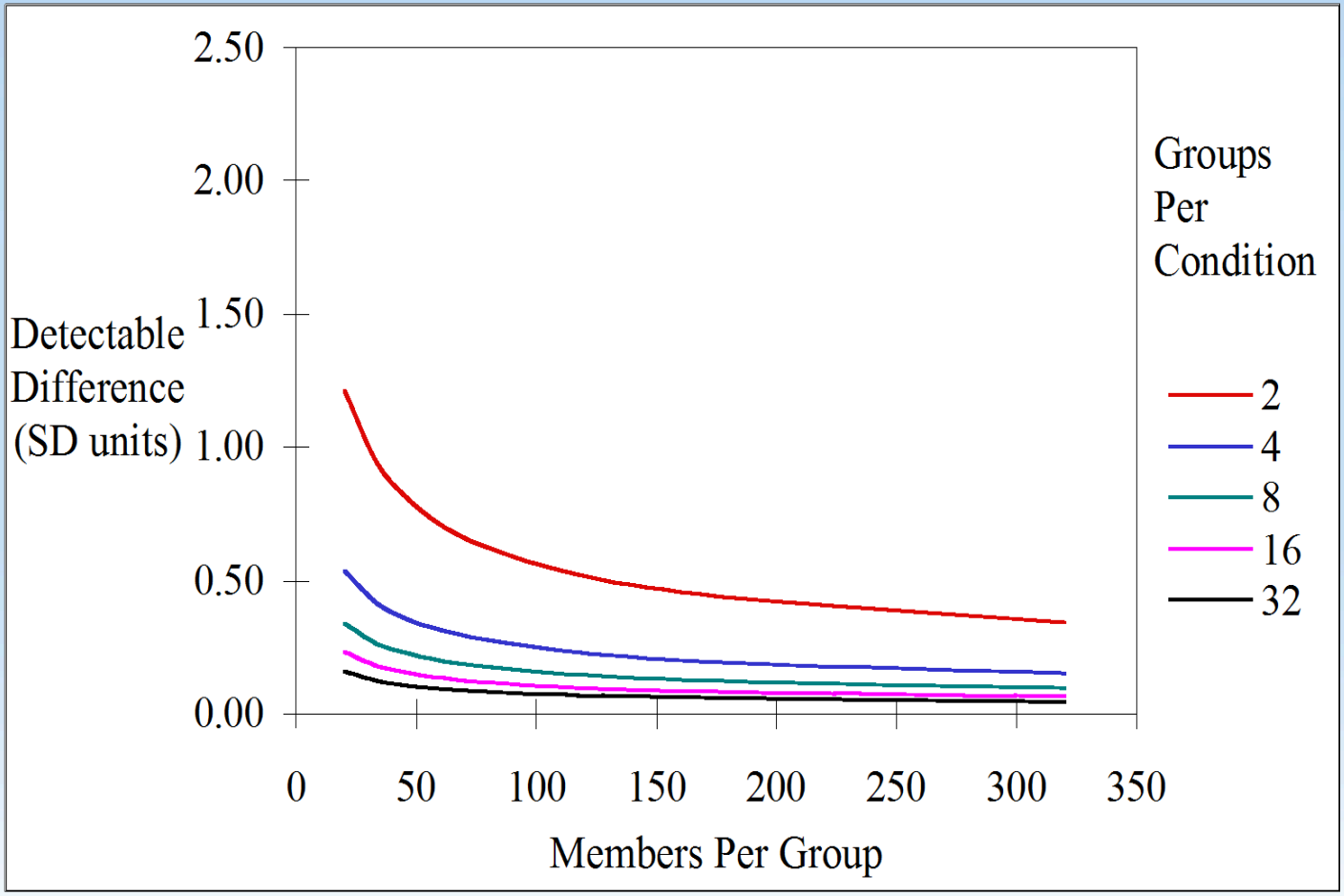
Detectable Difference

- 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)



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:

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

  - In a simple GRT, this expression becomes:

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

A Sample Size Example

- Calculate the required sample size per condition for a two-condition RCT, with 5% two-tailed Type I error rate and 80% power for a detectable difference of 0.2 standard deviations.

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

- To perform the calculations in standard deviations, set $\sigma_y^2 = 1$.
- Substitute this expression into the formula for the sample size to determine how many participants must be randomized to each condition.

$$m = \frac{2(1)(1.96+0.84)^2}{(0.2)^2} = 392$$

A Sample Size Example

- Calculate the required sample size per condition for a two-condition GRT, with 5% two-tailed Type I error rate and 80% power for a detectable difference of 0.2 standard deviations, given an ICC estimate of 0.01 and 100 members per group.

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

$$g = \frac{2(1)\left(1+\left(m-1\right)0.01\right)\left(1.96+0.84\right)^2}{100\left(0.2\right)^2} = 7.8$$

A Sample Size Example

- We cannot stop at this point, because the critical values for t used in this calculation are not matched to the df calculated using the result.
- df=2(g=1)=2(8-1)=14.
- The critical values for t based on 14 df are 2.145 and 0.868.
- We repeat the calculation using those values.

$$g=\frac{2(1)(1+(m-1)\,0.01)(2.145+0.868)^2}{100(0.2)^2}=9.03$$

A Sample Size Example

- df=2(g=1)=2(9-1)=16.
- The critical values for t based on 16 df are 2.12 and 0.865.

$$g=\frac{2(1)\left(1+(m-1)\,0.01\right)(2.12+0.865)^2}{100(0.2)^2}=8.86$$

- We can stop at this point, as the result matches the value used to calculate the critical values for t.
- There will be 80% power for a two-tailed Type I error rate of 5% to detect a 0.2 sd effect given an ICC of 0.01 and m=100 with 9 groups per condition.
- It would be wise to perform a sensitivity analysis using several values of the ICC and m if those estimates may vary.

Unbalanced Designs

- As long as the ratio of the largest to the smallest group is no worse than about 2:1, the methods presented above are fine.
- Given more extreme imbalance, other methods are required.
  - For a GRT, several recent sources provide alternative methods.
    - van Breukelen G, Candel M, Berger M. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. Statistics in Medicine. 2007;26(13):2589-603.
    - Candel MJ, Van Breukelen GJ. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. Statistics in Medicine. 2010;29(14):1488-501.
    - You Z, Williams OD, Aban I, Kabagambe EK, Tiwari HK, Cutter G. Relative efficiency and sample size for cluster randomized trials with variable cluster sizes. Clinical Trials. 2011;8(1):27-36.
    - Candel MJ, Van Breukelen GJ. Repairing the efficiency loss due to varying cluster sizes in two-level two-armed randomized trials with heterogeneous clustering. Statistics in Medicine. 2016;35(12):2000-15.
    - Moerbeek M, Teerenstra S. Power analysis of trials with multilevel data. Boca Raton: CRC Press; 2016.

Unbalanced Designs

- As long as the ratio of the largest to the smallest group is no worse than about 2:1, the methods presented above are fine.
- Given more extreme imbalance, other methods are required.
  - For an IRGT, see
    - Candel MJ, Van Breukelen GJ. Varying cluster sizes in trials with clusters in one treatment arm: sample size adjustments when testing treatment effects with linear mixed models. Statistics in Medicine. 2009;28(18):2307-24.
    - Moerbeek M, Teerenstra S. Power analysis of trials with multilevel data. Boca Raton: CRC Press; 2016.

Summary

- The usual methods for detectable difference, sample size, and power must be adapted to reflect the nested design.
- Power for GRTs and IRGTs is tricky, and investigators are encouraged to collaborate with a biostatistician.
- Both of Cornfield's penalties must be addressed: extra variation and limited df.
- Failure to do so will result in an inflated Type I error.
- There are effective design and analytic methods to reduce the extra variation.
- The most important factors affecting power in a GRT are the ICC and the number of groups per condition.
- Investigators should seek good estimates for those parameters.

Pragmatic and Group-Randomized Trials in Public Health and Medicine

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

- Provide feedback on this series
- Download the slides, references, and suggested activities
- View this module again
- View the next module in this series:
  Part 5: Examples

Send questions to:

GRT@mail.nih.gov

National Institutes of Health
Office of Disease Prevention