Book Review

Power Analysis of Trials with Multilevel Data. M. Moerbeek and S. Teerenstra (2016). Boca Raton, FL: Chapman & Hall / CRC Press. 288 Pages, ISBN: 9781498729895.

I enjoyed reviewing the new CRC Press / Chapman Hall book entitled *Power Analysis of Trials with Multilevel Data*, by Mirjam Moerbeek and Steven Teerenstra. This book addresses a critical need in the scientific community for a well-organized, easily accessible guide to performing power analysis and computing required sample sizes for randomized trials embedded in multilevel study designs, where observations of interest are nested within higher-level units (e.g., patients within clinics or repeated measures on participants). Multilevel study designs introduce dependencies in the collected data that can substantially reduce effective sample sizes (e.g., patients coming from the same clinic tend to have correlated outcomes), and these dependencies need to be correctly accounted for at the sample size calculation stage of a study design. Unfortunately, many applied researchers fail to do this when performing power and sample size calculations, and this can limit the realized statistical power of otherwise well-designed studies. This book effectively compiles all the published literature on this specialized topic, putting it in one place for researchers who design these types of studies and could benefit from a concise and practical resource on this important aspect of study design. The two Dutch authors are experts in this area and are very well-equipped to provide more general education and practical advice on this topic. Multilevel study designs in which power analysis methods for independent observations do not apply are quite common, but no prior books have attempted to organize all the possible power analysis approaches for these types of studies into a single reference.

The first three chapters provide important background for the later chapters on specific types of study designs. Chapter 1 presents an overview of designing randomized trials in multilevel contexts and addresses the importance of representative multistage sampling procedures for appropriate population inferences, in addition to the fact that these designs produce hierarchical data structures. This chapter is an overview (and not a “how-to” guide) on designing randomized trials. The authors provide basic guidelines for choosing between different randomized designs and emphasize the importance of careful sample size calculation for cost efficiency. Chapter 2 continues with a basic overview of multilevel modeling, with several examples. I would have liked to see some coverage of appropriate methods for testing hypotheses about the variances of random coefficients in multilevel models with multiple random effects, but this is a minor quibble. A larger issue that I had was the focus in this chapter (and throughout the book) on continuous and binary outcomes only; some trials focus on more general categorical outcomes or count outcomes, and alternative models are needed in these cases. In addition, reference to more recent work by Kim et al. (2013) would have been helpful in assisting readers with choices of estimation methods when fitting generalized linear mixed models to binary outcomes. Finally, nlmer is not a package in R; the package to which I believe the authors were referring is lme4, where readers can find the lmer and glmer functions for fitting these types of models. Chapter 2 can be skipped by readers familiar with multilevel modeling but would serve as a good introduction for readers not very familiar with this topic.

Chapter 3 presents a nice overview of many common types of power analyses. The focus is mainly on simple comparisons of means and proportions, which are common in randomized trials, and useful formulas are defined and described. However, power analyses for more complicated types of objectives outside of testing simple treatment effects (e.g., testing interactions) are not covered. I do wish that there had been more discussion of power analysis approaches when *treatment effect heterogeneity* is of interest, and researchers wish to power a study to detect interaction effects of a specified size between a treatment factor and other hypothesized effect moderators (e.g., Demidenko 2008; Imai and Strauss 2011; Tipton 2013). The authors do include a useful discussion of simulation-based approaches, and I liked the fact that they emphasized the importance of reserving enough time for careful power analyses (and not throwing something together two days before a grant proposal is due, which is a common nightmare for statistical consultants). The authors provide a very practical, step-by-step procedure for performing careful power analyses, including an example, and many readers will find these steps helpful, especially statistical consultants. They also provide an overview of past studies that have considered sample size calculations in the multilevel context or used simulation studies to develop guidelines in this context, motivating the more detailed discussions in later chapters.

Chapter 4 is the first "application" chapter, presenting approaches for cluster randomized trials (CRTs). The authors discuss why these types of studies, despite their popularity, result in losses in statistical efficiency. This chapter introduces the basic structure that the authors follow in all their "application" chapters (4 through 10): they start with a nice example of the multilevel model that one would generally fit to the collected data in this setting (and continuously emphasize that many researchers still do not employ these approaches in multilevel studies, when they most certainly should be doing so); they then turn to a discussion of the key factors affecting power in this particular context; next they provide formulas for computing required sample sizes at all levels of the design, emphasizing the roles that effect sizes, intra-class correlations (ICCs), and other key parameters play in the calculations, depending on whether comparisons of means or comparisons of proportions are of interest; and, in one of the finest contributions of this book (in my view), they stress how *cost constraints* should be incorporated into the sample size calculations in a particular context. They then end the chapter with a nice, fully-worked example of how to apply the calculations in practice, clearly describing and motivating all the design decisions. I found this to be a very effective and practical structure for presenting this rather complex material. An additional minor quibble that I had is that it would have been nice to incorporate the software discussed in the final chapter (Chapter 12) into the (very helpful) examples at the end of each "application" chapter, clearly illustrating how to use the available software to perform the various calculations given the design objectives and input parameters.

Chapter 5 turns to how to improve power in CRTs, mainly via the inclusion of cluster-level covariates in the multilevel models (reducing components of variance). The authors discuss design strategies for improving the balance of cluster-level covariates across treatment conditions, such as stratified randomization and the utility of repeated measures designs for increasing power. They also discuss various types of *crossover designs* in the hierarchical context and their increased efficiency relative to CRTs. The authors once again explicitly show how costs should be considered in these cases, which is a great contribution. They add a nice discussion of *stepped wedge designs*, which aren't often discussed in this context. This adds to the comprehensiveness of the volume from a design point-of-view.

Chapter 6 covers *multisite* trials, where randomization is done at the subject level within a site (e.g., clinic), but again analysts often do not account for clustering by site. I especially liked how the authors demonstrated the calculation of an ICC when a treatment effect can randomly vary across sites and what role this ICC plays in the sample size calculations. As in all the other application chapters, the authors carefully weigh the pros and cons of these types of designs, liberally including citations to published literature to support their various assertions. Chapter 7 turns to *pseudo cluster randomized trials*, where randomization of a treatment occurs at the cluster level first, and then the majority (but not all) of the subjects within a cluster is randomized to that treatment. Given the complexity of the calculations introduced by these types of designs, incorporating the software into the worked example in Chapter 7 would have been especially helpful. Chapter 8 discusses the case of randomization at the subject level and then where nesting within groups occurs *after* the assignment (e.g., group therapy or multiple subjects visiting a health professional). The authors again present methods for computing ICCs unique to these types of designs, which is an important contribution.

Chapter 9 turns to longitudinal intervention studies, where repeated measures of outcomes are collected on subjects that have been randomized to one of two treatments, and two-level models will ultimately be employed in the analysis. The authors explicitly state that their methods assume ignorable missing-at-random (MAR) mechanisms in this context. Following the structure mentioned above for Chapter 4, readers can clearly determine how many subjects to study in each group and how many repeated measures should be collected on each subject to meet design objectives (including cost constraints). Importantly, the authors also extensively discuss the effects of dropout on power calculations. They provide a very nice example, emphasizing the importance of accounting for time-varying covariates in the models of interest to reduce within-individual variance (and thus increase power) if the computed sample size seems to be excessively large.

Chapter 10 presents calculations for more advanced designs, including designs with three levels of nesting and factorial designs with multiple treatments. Formulas arising in these more complicated contexts are clearly presented and described, and the authors continue to focus on cost constraints. They also present methods for designs collecting repeated measures in CRTs, which is a tremendously useful contribution. The coverage of factorial designs is brief, and no example is provided in this context (unfortunately). Chapter 11 presents a nice discussion of what to do when ICCs are not known, given the importance of ICCs for all the sample size calculations presented in the book. They discuss various approaches for estimating ICCs, such as using estimates from prior research, sample size re-estimation under different scenarios, Bayesian approaches, and (briefly) maximin optimal designs. Chapter 11 includes one of the true "golden nuggets" in this book: a comprehensive, four-page table (Table 11.1) of published studies presenting estimates of ICCs in a wide variety of different contexts (different fields, different subjects, different types of clusters); this was nothing short of a fantastic contribution. The book concludes with an overview of available software for performing these sample size calculations in Chapter 12, including new software (SPA-ML) developed by the authors that requires a MATLAB compiler runtime library (MCR) to be installed on a Windows machine. The authors provide a web link for obtaining this freeware and include screen shots illustrating the basic features of the software. They also link the various choices of input parameters back to the specific equations presented in the book, which is certainly useful. Again, it would have been nice to see this new software applied in at least a couple of the end-of-chapter examples, especially for designs requiring more complex calculations.

In sum, this will be a very useful book for researchers, statisticians, and consultants responsible for designing various types of randomized trials in multilevel settings. My minor quibbles are far outweighed by the important contributions that this single resource on power analysis in multilevel designs will make to the scientific community.

REFERENCES

Demidenko, E. (2008). Sample size and optimal design for logistic regression with binary interaction. *Statistics in Medicine*, 27, 36-46.

Imai, K. and Strauss, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, 19(1), 1–19.

Kim, Y., Choi, Y., and Emery, S. (2013). Logistic Regression with Multiple Random Effects: A Simulation Study of Estimation Methods and Statistical Packages. *The American Statistician*, 67(3), 171-182.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38(3), 239–266.

Brady T. West
Survey Research Center
Institute for Social Research
University of Michigan-Ann Arbor
bwest@umich.edu