References

[1] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289–300, 1995.

[3] Y. Benjamini and D. Yekutieli. The control of the false discovery rate: A practical and powerful approach to multiple testing under dependency. *Annals of Statistics*, 29:1165–1188, 2001.

[4] H. Bang, S. Jung, and S.L. George. Sample size calculation for simulation-based mulitple-testing procedures. *Journal of Biopharmaceutical Statistics*, 15:957–967, 2005.

[5] Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300, 1982.

[6] Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersectionunion tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996.

[7] Howard S. Bloom. The core analytics of randomized experiments for social research, 2006.

[8] F Bretz, T. Hothorn, and P. Westfall. *Multiple Comparisons Using R*. Chapman & Hall/CRC, 2010.

[9] J. Chen, J. Luo, K. Liu, and D. Mehrotra. On power and sample size computation for multiple testing procedures. *Computational Statistics and Data Analysis*, 55:110–122, 2011.

[10] Xutao Deng, Jun Xu, and Charles Wang. Improving the power for detecting overlapping genes from multiple dna microarray-derived gene lists. *BMC Bioinformatics*, 9, 2008.

[11] Nianbo Dong and Rebecca Maynard. Powerup!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, 6(1):24–67, 2013.

[12] S. Dudoit, J.P. Shaffer, and J.C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

[13] Olive Jean Dunn. Estimation of the medians for dependent variables. *The Annals of Mathematical Statistics*, 30(1):192–197, 1959.

[14] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[15] Y. Ge, S. Dudoit, and T.P. Speed. Resampling-based multiple testing for microarray data analsysis. *Test*, 12:1–77, 2003.

[16] Larry V. Hedges and Christopher Rhoads. Statistical power analysis in education research. Technical report, National Center for Special Education Research, 2010.

[17] S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6(2):65–70, 1979.

[18] W. Maurer and B. Mellein. *On New Multiple Test Procedures Based on Independent P-values and the Assessment of their Powers*, pages 48–66. Springer, 1988.

[19] Kristin E. Porter. Statistical power in evaluations that investigate effects on multiple outcomes: A guide for researchers. *Journal of Research on Educational Effectiveness*, 11:267–295, 2018.

[20] P.H. Ramsey. Power differences between pairwise multiple comparisons. *Journal of American Statistical Association*, 75:479–487, 1978.

[21] S.W. Raudenbush, J. Spybrook, R. Congdon, X. Liu, A. Martinez, H. Bloom, and C Hill. Optimal design plus empirical evidence (version 3.0). Technical report, 2011.

[22] Peter Z. Schochet. Guidelines for multiple testing in impact evaluations of educational interventions. final report. Technical report, Mathematica Policy Research, Inc. P.O. Box 2393, Princeton, NJ 08543-2393, 2008.

[23] Stephen Senn and Frank Bretz. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*, 6:161–170, 2007.

[24] Juliet Popper Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584, 1995.

[25] Jessica Spybrook, H.S. Bloom, Richard Congdon, Carolyn J. Hill, Andres Martinez, and Stephen W. Raudenbush. Optimal design plus empirical evidence: Documentation for the optimal design software version 3.0. Technical report, 2011.

[26] J.W. Tukey. The problem of multiple comparisions. Technical report, Princeton University, 1953.

[27] Peter H Westfall and S Stanley Young. *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*, volume 279. John Wiley & Sons, 1993.

[28] Peter H Westfall, R.D. Tobias, and R. D. Wolfinger. *Multiple Comparisons and Multiple Tests using SAS*. The SAS Institute, 2011.

[29] W. Corrin, S. Sepanik, R. Rosen, and A. Shane. Addressing early warning indicators: Interim impact findings from the investing in innovation (i3) evaluation of diplomas now, 2016.

[30] A. Gelman, J. Hill, and M. Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5:189–211, 2012.

[31] A. Gelman, J. Hill, and M. Yajima. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[32] Luke Miratrix, Michael Weiss, and Brit Henderson. An applied researchers guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. *Journal of Research on Educational Effectiveness*, 2020, forthcoming.