Rerandomization to Improve Baseline Balance in Educational Experiments

Kari Lock Morgan, Anna Saavedra, Amie Rapaport

Background / Context: Randomized experiments are the gold standard for making causal inferences, primarily because on average they balance baseline covariates across the treatment groups and hence provide unbiased impact estimates. However, balance on average (across many randomizations) does not guarantee balance for any particular randomization, and it's quite possible (or even likely, with many baseline variables) that some baseline variables may be imbalanced just by chance. Such baseline imbalance then creates conditional bias for the impact estimate. The dangers of relying on pure randomization for balance have been extensively noted (e.g. Gosset (1938), Seidenfeld (1981), Urbach (1985), Krause and Howard (2003), Rosenberger and Sverdlov (2008), Rubin (2008), Keele et al. (2009), Worrall (2010)). Typical practice involves checking for baseline equivalence after the experiment has been conducted, but by this point if imbalance is observed, additional analytic assumptions are required to correct for this imbalance. Rather than waiting until after the experiment has been conducted to address the problem, studies should instead prevent such imbalance as part of the experimental design. One flexible and intuitive way of doing this is to check balance after randomizing, and *rerandomize* if an allocation yields observed imbalance. This idea has been informally advocated by many (Sprott and Farewell (1993), Rubin (2008), Bruhn and McKenzie (2009), Cox (2009)), but requires a principled procedure to ensure that important properties such as unbiasedness and the nominal Type I error rate are maintained.

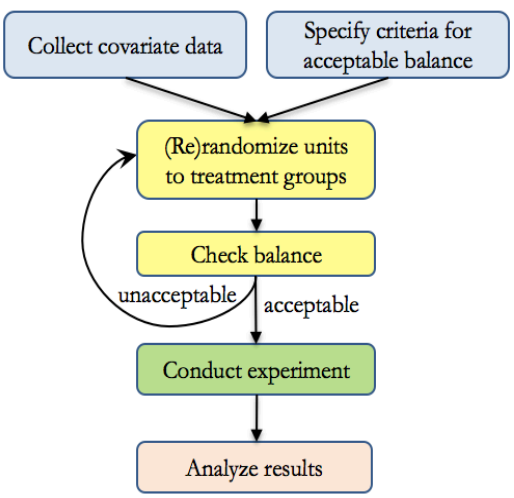

Figure 1. Rerandomization.

Research Design: A principled procedure for conducting rerandomization is described, depicted in Figure 1. Before randomization, criteria for acceptable balance must be specified based on baseline data available at the time of randomization. Units are then randomized into treatment groups, balance is checked, and units are rerandomized if the criteria for acceptable balance is not met. Only once an allocation meets the balance criteria are units actually assigned treatment. This research discusses conditions for maintaining unbiasedness, analytic procedures for maintaining valid inferential properties, and benefits for both baseline balance and precision of the estimated impact. Benefits will be derived theoretically for specific criteria and demonstrated empirically via a real educational example.

Findings / Results: Rerandomization maintains unbiasedness if every unit still has equal probability of being in either treatment group, which is most easily satisfied if the criteria for acceptable balance is objective and blind to group labels. Although rerandomization was designed to improve baseline balance, if variables balanced are correlated with the outcome, rerandomization will also increase the precision of the estimated treatment effect, and thus increase power if rerandomization is accounted for in the analysis. To take advantage of these gains in power, analysis must reflect the rerandomization procedure, for example by randomization-based inference. Randomization-based inference taking the rerandomization into account yields exact p-values and maintains the nominal Type I error rate, provided the criteria leaves enough acceptable randomizations for inference. Not accounting for rerandomization in analysis will still result in "valid" results in the sense that significant p-values can be trusted, the Type I error rate will no larger than as stated, and confidence intervals will have at least the nominal coverage, but results will be conservative, meaning that p-values could be smaller and intervals could be narrower if the rerandomization were taken into account. Mathematical derivations and theoretical results can be found in Morgan and Rubin (2012, 2015) for those interested, but the focus here is not to dwell on the mathematical derivations of balance improvement for specific criteria, but rather to convey the rerandomization method to an education audience and discuss its practical use in educational experiments. Therefore, the emphasis will be on the educational example described below.

Educational Example: Rerandomization was employed in the design of an experiment evaluating the impact of "Knowledge in Action (KIA)," a form of project-based learning, in Advanced Placement classes. Rerandomization was used in a cluster randomized trial to assign schools to the intervention or control groups across five large, urban school districts. The rerandomization criteria was slightly different in each district, due to differing baseline data availability, but each district imposed balance on some measure of baseline academic achievement measured by standardized test scores and socio-economic status. Randomization had to occur well before the start of the academic year, so rerandomization imposed balance on baseline variables measured on students from the previous year's class, rather than the current students. In Figure 2 we show, for one of the five districts, the resulting improvement in balance for these baseline variables. Zero represents perfect mean balance between treatment groups, and rerandomization yields a distribution of difference in means more closely concentrated around zero, eliminating imbalanced randomizations. The balance for the actual experiment is depicted with a black dot, and is very good for both variables, as enforced by rerandomization. The exact amount that this improvement in balance will increase precision will depend on the extent to which the baseline variables are correlated with each outcome, but for example, the resulting precision of the outcome difference in means would increase by a factor of about 2 for outcomes with $R^2 \approx 0.55$, an improvement equivalent to doubling the sample size!

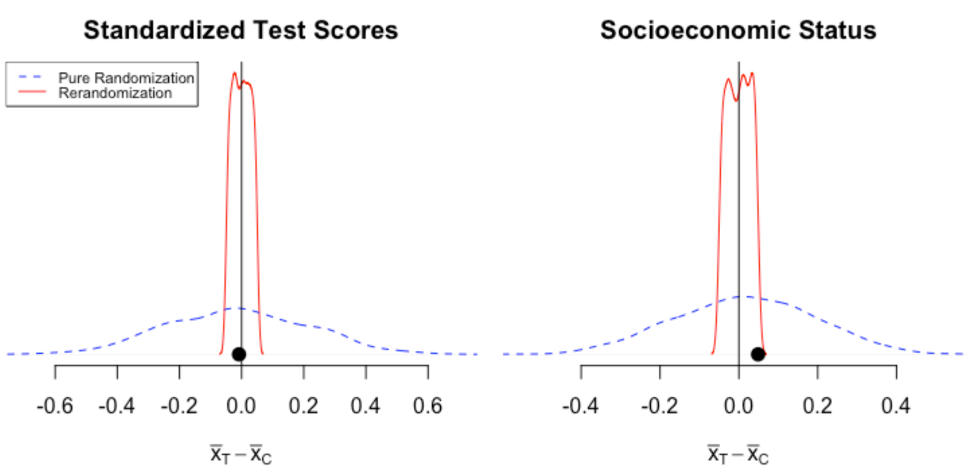

Figure 2: The difference in means across 1000 simulated pure randomizations ignoring baseline variables (in blue) and 1000 simulated rerandomizations enforcing baseline balance (in red).

Conclusions: While randomization ensure baseline balance on average, rerandomization preserves the "gold standard" benefits of randomization, but goes one step further and ensures baseline balance for your actual experiment, leading Tukey (1993) to call it the "platinum standard".

Employing rerandomization improves baseline balance and can increase power when the outcome is correlated with the baseline variables being balanced. This improves causal inferences by improving the chance of finding significant results when an impact exists, and lending credibility to significant results by helping to assure that any observed impact is really due to the intervention, not chance baseline imbalance.

References:

Bruhn, M. & McKenzie, D. (2009). "In pursuit of balance: Randomization in practice in development field experiments," *American Economic Journal: Applied Economics*, **1**: 200-232.

Cox, D.R. (2009). "Randomization in the Design of Experiments," *International Statistical Review*, **77**: 415-429.

Gosset, W.J. (1938). "Comparison between balanced and random arrangement of field plots," *Biometrika*, **29**: 363.

Keele, L., McConnaughy, C., White, I., List, P.M.E.M, & Bailey, D. (2009). "Adjusting experimental data," In *Experiments in Political Science Conference.*

Krause, M.S. & Howard, K.I. (2003) "What random assignment does and does not do," *Journal of Clinical Psychology,* **59**: 751-766.

Morgan, K.L & Rubin, D.B. (2012). "Rerandomization to Improve Covariate Balance in Experiments," *The Annals of Statistics,* **40**(2): 1262-1282.

Morgan, K.L. & Rubin, D.B. (2015). "Rerandomization to Balance Tiers of Covariates," *Journal of the American Statistical Association (JASA)*, **110**(512): 1412-1421.

Rosenberger, W.F. & Sverdlov, O. (2008). "Handling covariates in the design of clinical trials," *Statistical Science*, **23**: 404-419.

Rubin, D.B. (2008). "Comment: The design and analysis of gold standard randomized experiments," *Journal of the American Statistical Association (JASA)*, **103**: 1350-1353.

Seidenfeld, T. (1981). "Levi on the dogma of randomization in experiments," In *Henry E. Kyburg, Jr. & Isaac Levi* (R.J. Bogdan, ed.), 263-291, Springer, Berlin.

Sprott, D.A. & Farewell, V.T. (1993). "Randomization in experimental science," *Statistical Papers*, **34:** 89-94.

Tukey, J.W. (1993). "Tightening the Clinical Trial," *Controlled Clinical Trials,* **14**(4): 266-285.

Urbach, P. (1985). "Randomization and the design of experiments," *Philosophy of Science*, **52:** 256-273.

Worrall, J. (2010). "Evidence: Philosophy of science meets medicine," *Journal of Evaluation in Clinical Practice*, **16**: 356-362.