In Pursuit of Balance: Randomization in Practice in Development Field Experiments

Miriam Bruhn (World Bank)
David McKenzie (World Bank)

Motivation

- Randomized experiments are increasingly being used in development economics
  - NEUDC 2002: 4 experiments
  - NEUDC 2008: 20 experiments
- In many cases, Researchers have control over the actual randomization
- Sample sizes tend to be small (< 500)
  - Question of not just whether to randomize, but how to do so.

Key issues

- Randomization ensures treatment and control groups have identical characteristics *on average*
  - But in small samples, any given draw could give different average characteristics.
  - E.g. Suppose 30% of sample are female. Chance that percent female in two groups will differ by more than 10% is:
    - 38% in sample of 50
    - 27% in sample of 100
    - 9% in sample of 200
    - 2% in sample of 400.
- Achieving Balance Particularly important for
  - Variables that are correlated with outcomes of interest
  - Characteristics that are used in sub-group analysis

What does this paper do?

- Takes stock of how randomization being done in practice
  - Surveys of researchers, study of papers
  - New simulations
- Draws lessons for key questions facing researchers.

How to achieve balance?

Several methods (with baseline data)

- Single random draw
- Randomizing within groups
  - Stratification
  - Pair-wise matching
- Re-randomization
  - Big Stick
  - Minmax t-stat

Stratification

- Pick one, two, three or more characteristics that are thought to influence outcomes
- Create groups (strata) and randomize within these groups
  - Women over 40, Women under 40, Men over 40, Men under 40
- Stratification does not remove all imbalances for continuous variables
- No consensus in literature on how many variables to use (overstratification?)
- Also no consensus on whether to control for strata dummies in ex-post analysis (not including dummies may lead to conservative standard errors)

Pair-Wise Matching

- Calculate “distance” between each pair of individuals in terms of **several** characteristics that influence outcomes
- Individuals with the smallest distance make a match → randomize within each match
  - Greedy algorithm: Pick pair with smallest distance, re-calculate distances among remaining individuals, repeat
  - Assign matches such that total distance between all pairs is minimized
- Overall computationally and time intensive
  - With 300 observations, need to calculate about 45,000 distances

Big Stick Rule

- Take a single random draw
- Examine difference in means in **several** characteristics that influence outcomes
- If any difference is statistically significant at 5 percent level (or other), re-draw, repeat
- Guards against “unlucky” draws

Minmax T-Stat Method

- Take 1000 random draws
- Pick the random draw that has the minimum maximum t-stat on the difference in means in **several** characteristics
- Can use other rules
  - No t-stat bigger than 1
  - Average t-stat equal 0.5
  - R-squared instead of t-stat
- Ex-post analysis should control for variables used for checking balance

Summary of 18 papers

Stratification	13
Pure randomization	3
Matched pairs	2

- Details of the methods typically not provided, such as
  - Number of strata
  - Strata dummies included in ex-post analysis
  - Distance metric for matching
  - Re-randomization in case of large differences
  - Rule for re-randomization
  - Public vs. private randomization

Survey of experts

- Responses from 25 out of 35 researchers
- Median researcher has done 5 randomized experiments

% of researchers who have ever used

	Unweighted	Weighted**	+5 experiments
Single draw*	80	84	92
Matching	56	52	54
Big Stick	12	15	15
Minmax t-stat	24	45	38

* Possibly with stratification **Weighted by number of experiments

Public vs. private randomization

- Transparency is often an argument for choosing randomization
- But most randomizations are done in private
- Influences choice of method since not all methods can be done in public

Panel Data

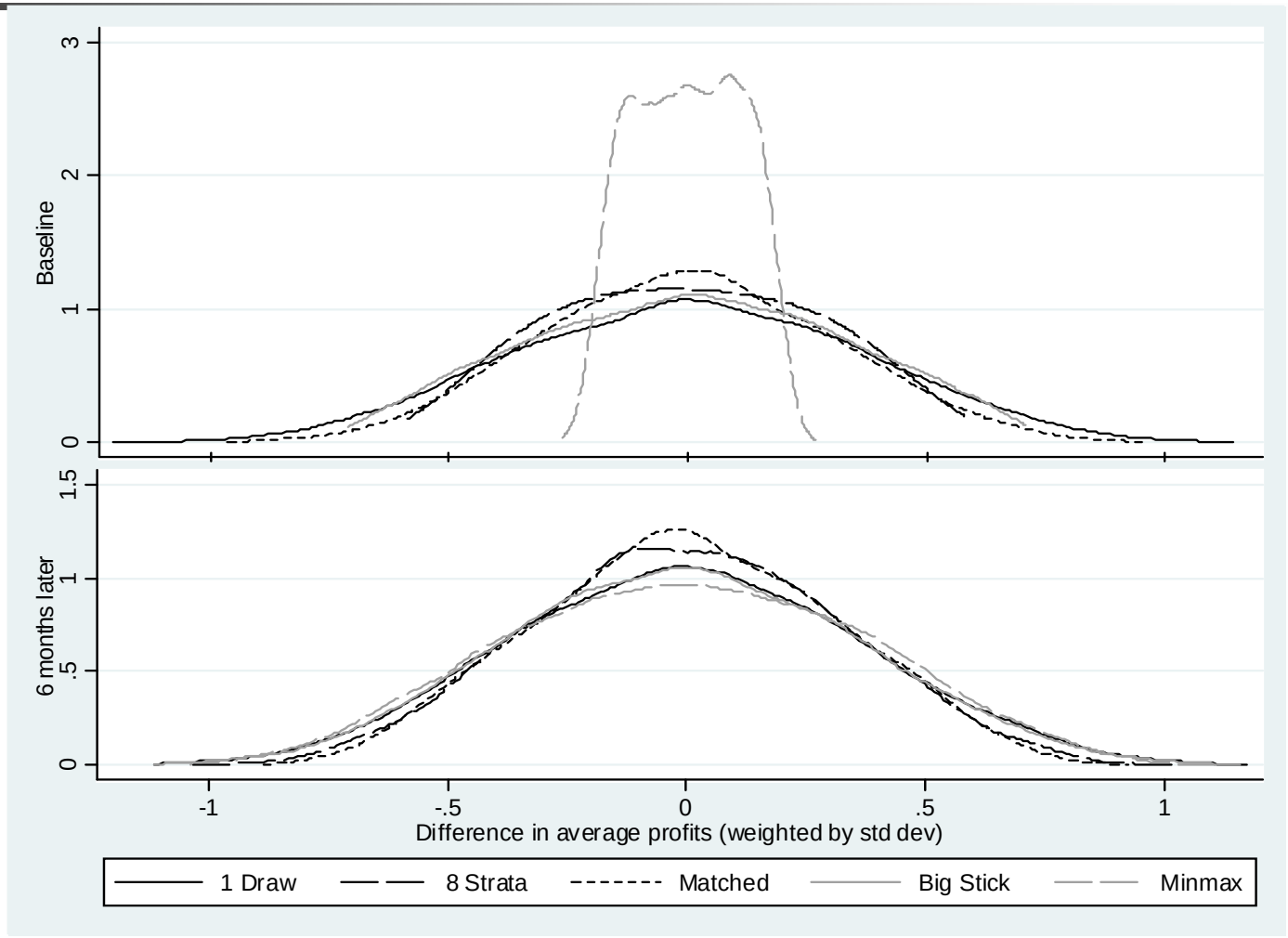
- Sir Lankan Microenterprises (de Mel et al)
  - Enterprise profits
- ENE (Mexican Labor Market Survey)
  - Income
- IFLS (Indonesian Family Life Survey)
  - School attendance, household expenditure
- LEAPS project Pakistan
  - Math test scores and height z-scores.

=> For each, simulate assignments to treatments, and also treatment.

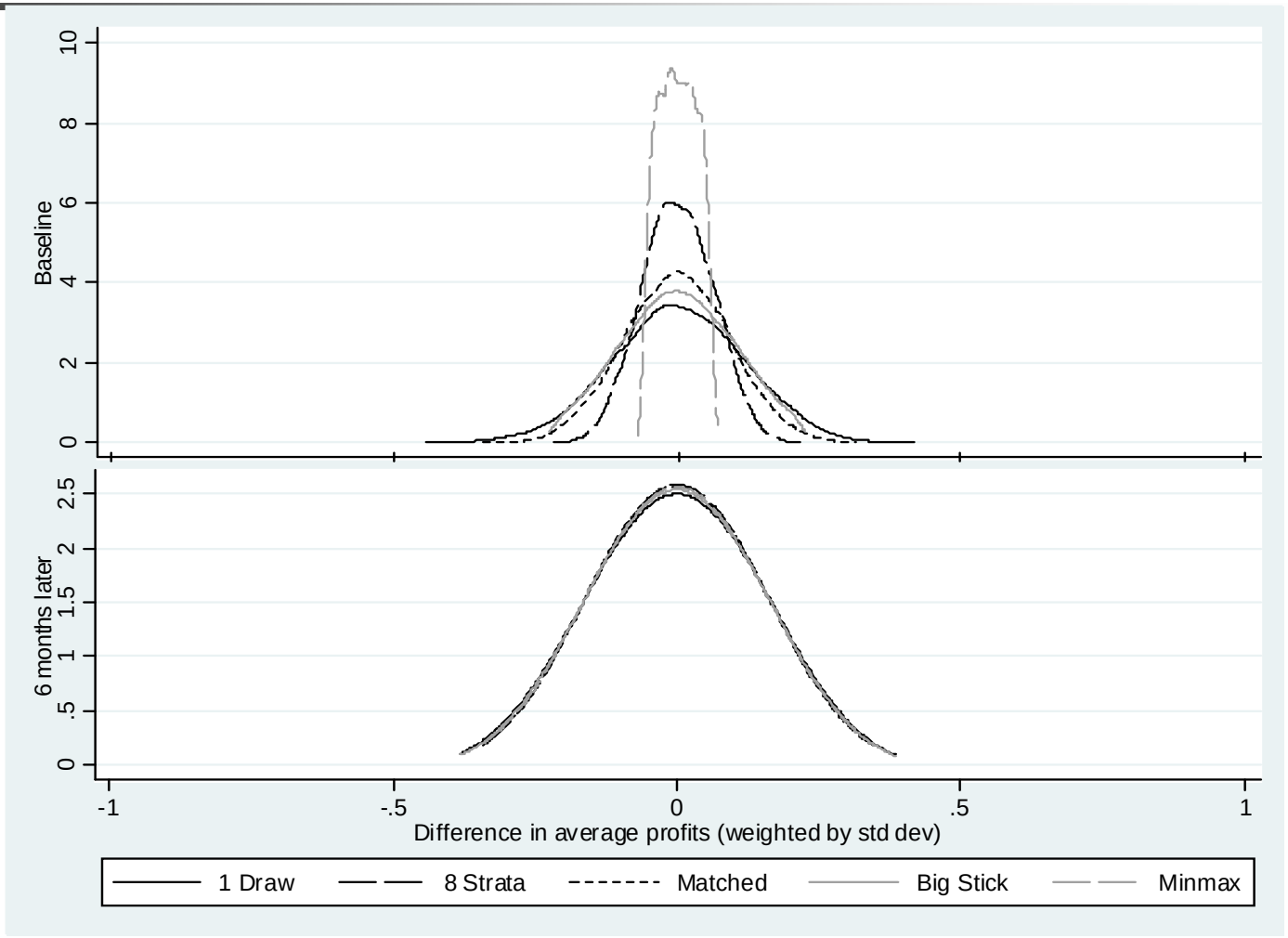
Simulations

- Subsamples of 30, 100, and 300 observations
- 10,000 bootstrap iterations
- 5 different methods
  - Single random draw
  - Stratification based on 2 variables (8 strata), 3 variables (24 strata), 4 variables (48 strata)
  - Pair-wise greedy matching (Mahalanobis distance)
  - Big stick method: re-draw if any difference is significant at less than 5 percent level
  - Picking draw with minimum maximum t-stat out of 1,000 draws

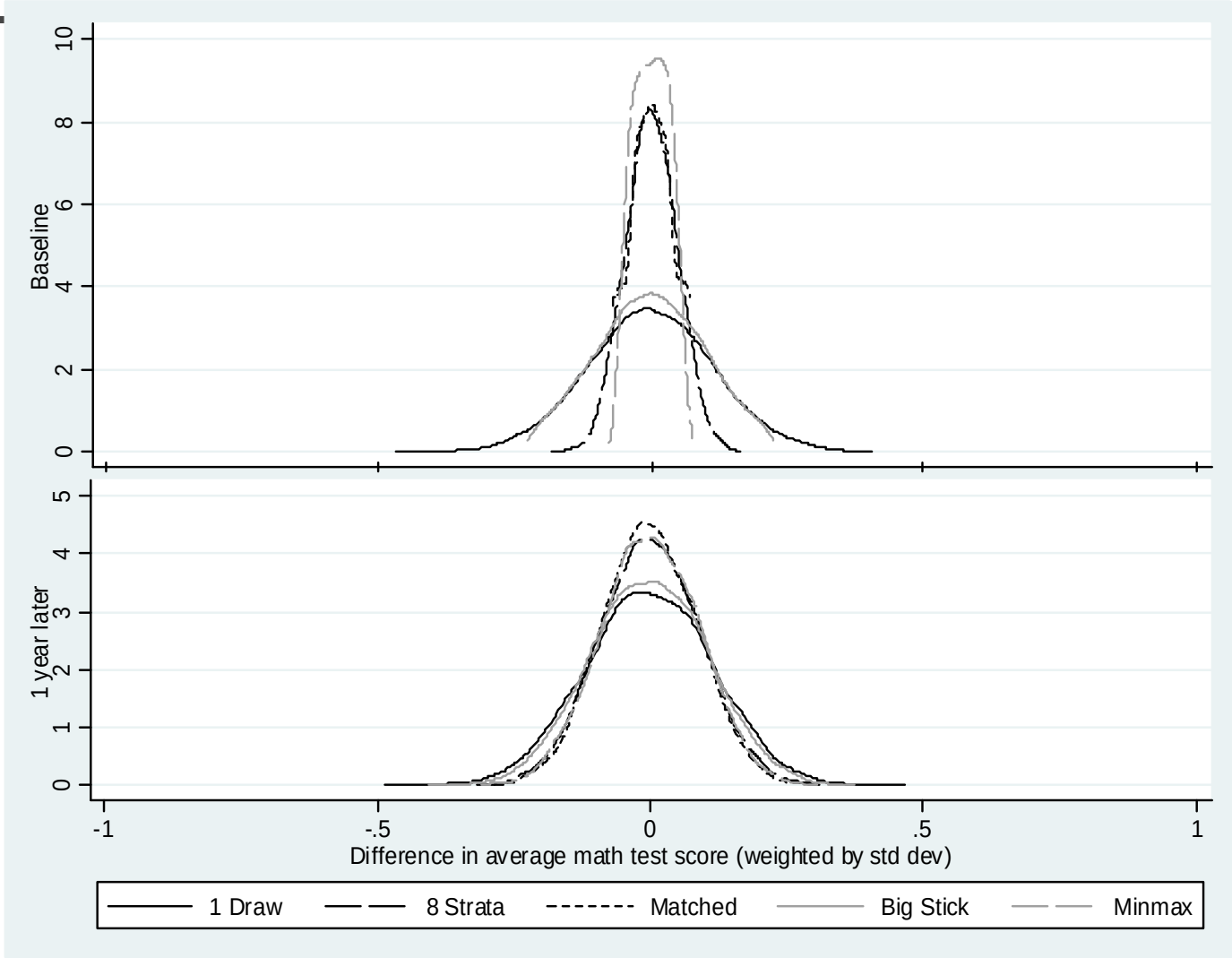
Differences: Sri Lankan Profit Data 30 observations



Differences: Sri Lankan Profit Data 300 observations



Differences: LEAPS math test scores: 300 obs



Which does better in terms of achieving balance and avoiding extremes?

- For variables like incomes and profits, which are not very persistent, all methods perform similarly, especially when sample sizes 100+
- For more persistent variables and smaller samples, pairwise matching does best, followed by stratification and re-randomization.

Do we need to control for the method?

- There is some disagreement about this issue, particularly for stratification
- Bottom line:
  - Our results show do need to control for method
    - Otherwise Size of Tests will be wrong
    - While on average is overly conservative not to include strata dummies, not necessarily the case – may give too small standard errors for any given draw.
    - Failure to control for method can result in less power than if a simple random draw was used.

How should inference be done after re-randomizing?

- Theoretically methods not clear
- Randomization inference/permutation tests statistically valid, but very messy to perform
- We recommend controlling for all variables used to check balance as regressors
  - Simulations show this seems to work in practice
- However, given that re-randomization offers no improvement in performance over matching and has more troublesome inference, we suggest researchers rethink using such methods.

What is the meaning of the standard Table 1?

- Most papers have Table testing for difference in means between treatment and control group.
- Since treatment is random, this assesses the probability that something occurred by chance when WE KNOW it occurred by chance.
- Statistical imbalance is immaterial when considering whether any difference between groups affects results
  - Control for differences in variables thought to affect outcome of interest, regardless of statistical significance.
- Other point to note is that if re-randomization done, degree of balance in this Table is overstatement of degree of balance achieved in other variables.

Recommendations

- Better reporting of randomization method is needed
  - a. Which randomization method was used and why?
  - b. Which variables were used for balancing?
  - c. For stratification, how many strata were used?
  - d. For re-randomization, which cutoff rules were used?
  - e. Who performed the randomization?
  - f. Was it public or private?
  - g. Done by computer, or manual randomization device?

Recommendations

- Re-think the common use of re-randomization – pairwise-matching performs as well or better
  - But some situations where can't form pairs.
- Stratify or match more on baseline outcome variable
- Take account of randomization method in analysis – control for strata or pair dummies.
- Choice of whether or not to control for a variable ex-post should not be driven by whether the baseline difference is statistically significant