Adaptive Trials in the Social Sciences: Prospects and Pitfalls

Don Green

Columbia University

Prepared for Presentation at the RCTs in the Social Sciences Conference

York, UK 2018

Outline

- Definition of “adaptive” design
- Advantages and disadvantages: Implications for analysis, planning, and ethics
- Connection to “multi-arm bandits”
- Simulations
- Empirical applications to politics

How is an adaptive design different from a static design?

- A static design applies the same procedures for allocating treatments and measuring outcomes throughout the trial
- An adaptive design may change the allocation of subjects to treatment arms based on interim results
  - Other adaptations: adjust the allocation of resources to different outcome measures or research sites…or change the treatment itself

Examples of Adaptive RCTs

- Example of changing allocation of subjects after certain treatment arms are declared inferior
  - Haynes et al. (2013) *Journal of Policy Analysis and Management*
  - Sykes et al. (2012): Systemic Therapy for Advanced or Metastatic Prostate Cancer: Evaluation of Drug Efficacy: five arms vs. control in 100 international centers. Two treatment arms were discontinued after initial results

Adaptations have a shady reputation

- Unsavory history of midcourse changes in design, some of which are poorly documented or guided by tendentious decision rules (e.g., ‘keep going until significance is achieved’)
- Post hoc changes in design make honest statistical inference challenging
- Growing concern about the corrosive role of discretion has led to calls for pre-analysis plans and standard operating procedures

What are the potential advantages of adaptive allocation of subjects in multi-arm trials?

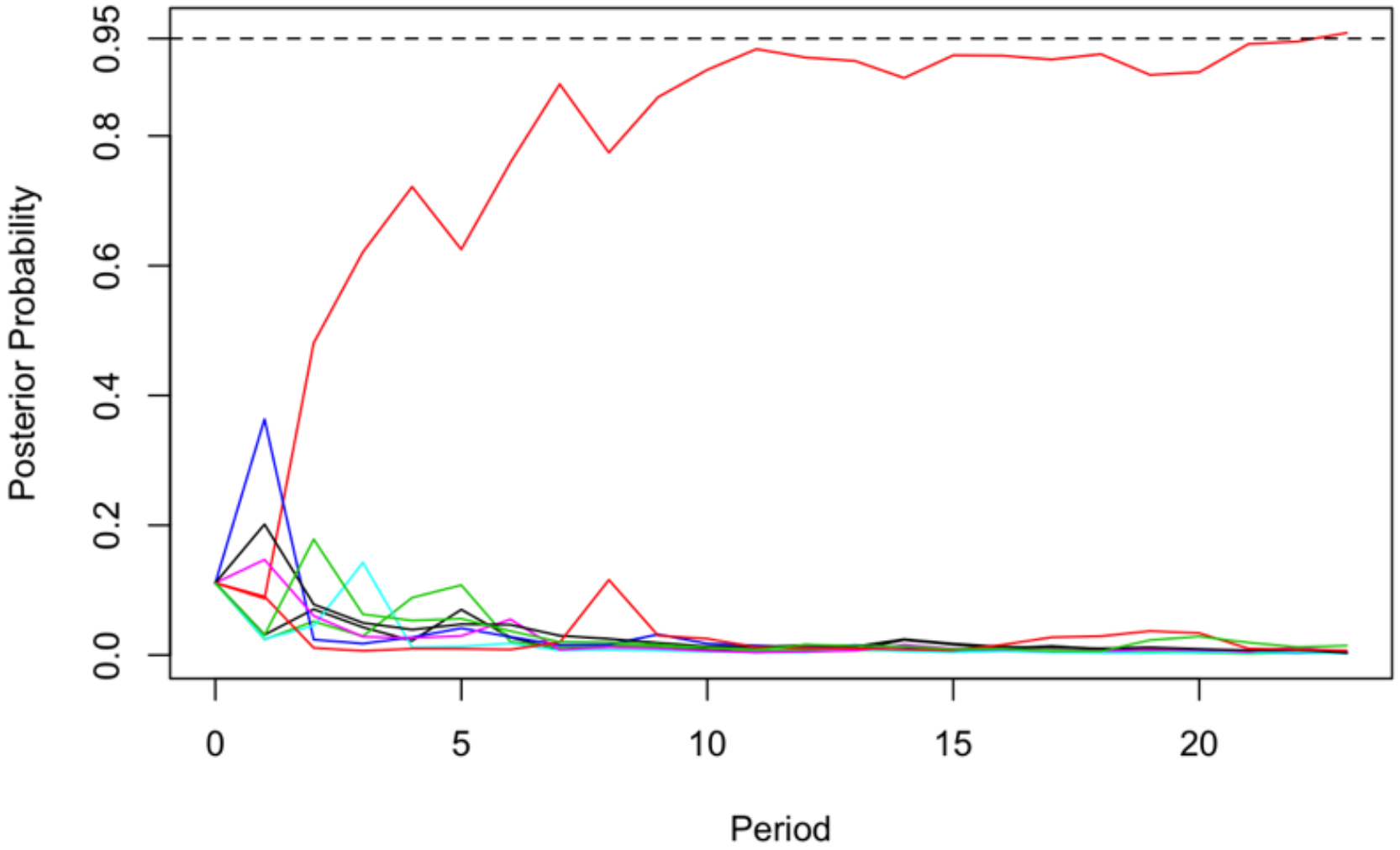
- Has the potential to discern the best-performing arm more quickly (fewer data-collection sessions, and fewer subjects) and more precisely
- Reduces ethical downside associated with allocating subjects to inferior treatment arms

Simulation of how adaptive designs may save resources and time

- Hypothetical RCT involving a control group and eight treatment arms
- Administer treatment and gather 100 outcomes during each “period”
- Each subject’s outcome is binary
- Allocate next period’s subjects according to “posterior” probabilities that a given treatment arm is best (see below), and stop when one arm reaches 0.95

How many periods go by before one arm emerges with a 95% chance of being the best?

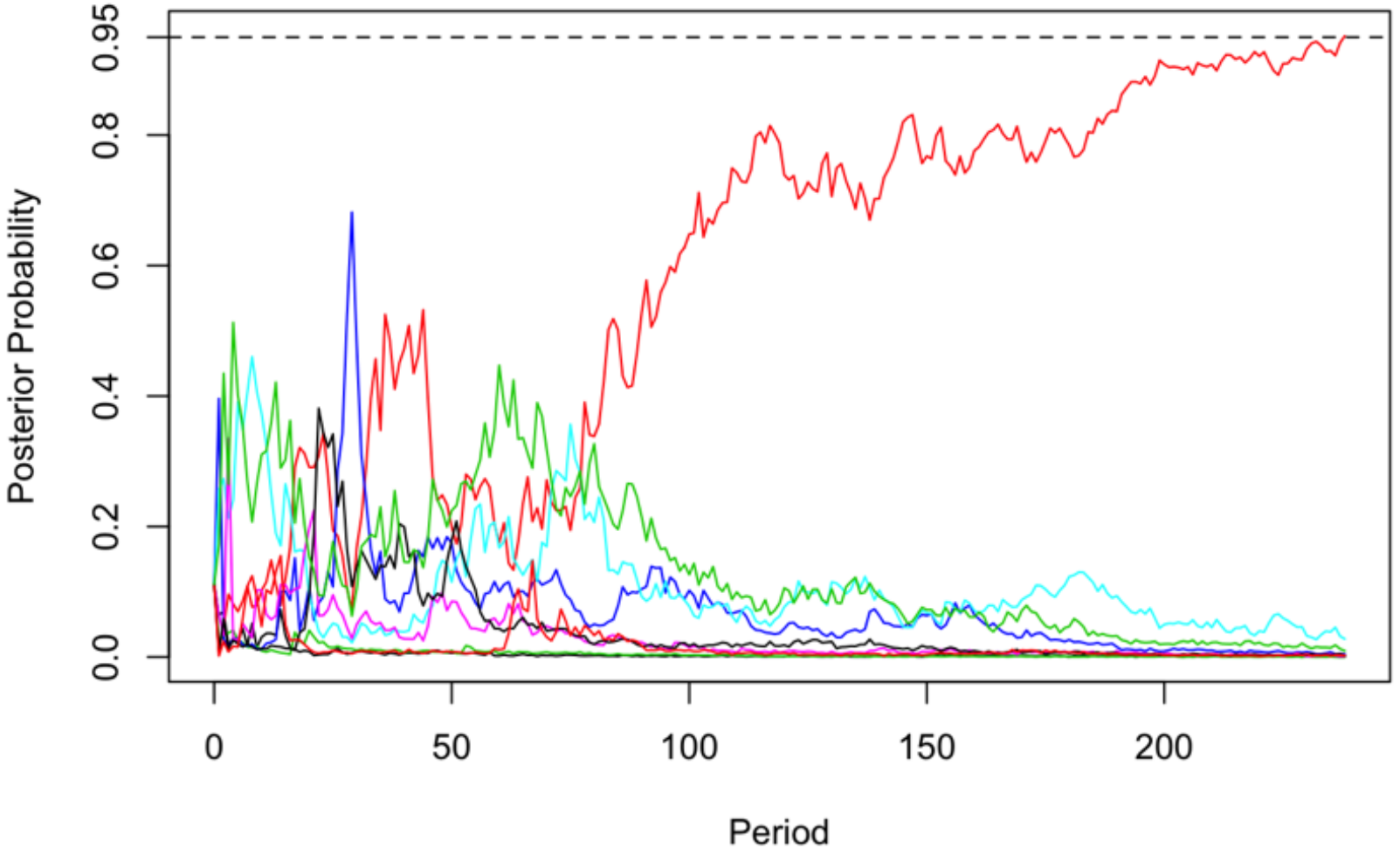
Scenario 1: Pr(success|Best) = .2 and .1 for other arms



What are the potential disadvantages?

- No guarantee that adaptive design will be superior in terms of speed or accuracy – e.g., in situations where there are many arms, all of which are equally (in)effective. May lead to a drawn out and futile search
- Even when there exists a truly a superior arm, the method may generate a long right tail of expended resources if by chance it gets off to a bad start.

Scenario 2: Pr(success|Best) = .12 and .1 for other arms
Many periods go by before one arm wins



Another disadvantage: bias

- Risk of bias in estimation of the properties of the winning arm due to the “winner’s curse” – lucky draws contribute to the winning arm’s victory
- Bias diminishes as the N allocated to the winning arm increases

Yet another disadvantage: managing implementation and analysis

- Increased complexity of conducting appropriate hypothesis tests and power calculations
- And if your focus is on the contrast between treatment and control, bear in mind that you may want a “static” control group (otherwise, it may garner few subjects)

What kinds of experiments lend themselves to adaptive design?

- Requires multiple periods of treatment and outcome assessment. (Won't work for voter turnout, where everyone's outcome is measured at the same time.)
- Well suited to survey, on-line, and lab experiments, where participants are treated/measured in batches over time. Some field experiments are conducted in stages, although the logistics of changing treatment allocations may be daunting.

Connection to optimization problems such as the “multi-arm bandit”

- Optimization problem of finding the best treatment arm.

- The MAB problem is framed as a trade off between learning about the relative merits of the various arms (exploration) and reaping the benefits of employing the best arm (exploitation). Not solely focused on estimating treatment effects.

How does the trade-off between “exploration” and “exploitation” affect the design of an RCT?

- Need some exploration or risk mistakenly settling on an inferior arm.
- Exploration also allows the researcher to adjust for time trends if the adaptive trial changes the assignment probability over time. Time trends might otherwise be correlated with treatment in ways that might jeopardize the interpretation of the results.
- Too much exploration will fail to capitalize on the apparent superiority of the best arm. A static design is an extreme case of an exploration-only design.

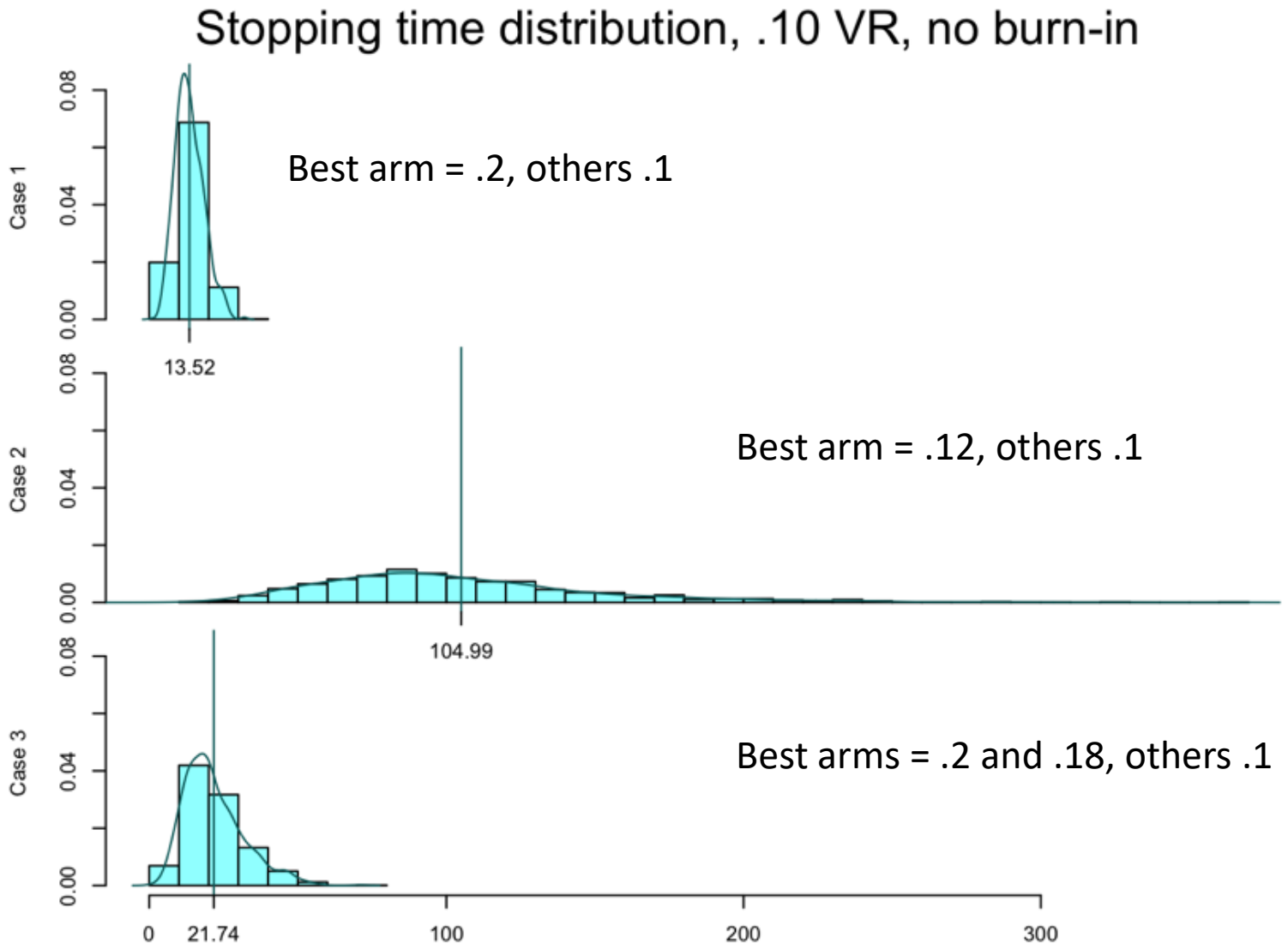
What are some widely used algorithms for automating “adaptation”?

- Methods such as “Thompson sampling” assess interim results periodically and allocate subjects to treatment arms in proportion to the posterior probability at that point that a given arm is best.
- The more likely an arm is to be “best,” the more subjects it receives.
- Many variants on this idea and no guarantees that any adaptive allocation method will improve on a static design

Adaptive designs require some forethought about stopping rules

- Fixed number of periods?
- Stop time based on whether the best arm passes some posterior probability threshold?
- May wish to focus on “value remaining” as a stopping rule to lessen the risk of futile search: the upper end of the 95% confidence interval of some other arm is no more than X% better than the apparent best arm

Three scenarios using a 10% “value remaining” stopping rule



What implications do adaptive designs have for analysis?

- One may think of a period during which a new array of assignment probabilities as a “block”
- Pooling over blocks may be done using inverse-probability weights (which in turn requires at least some allocation to every arm in every period) to avoid a confound between time and assignment probabilities and/or excessive weights
- Alternatives are to model over-time drift or to sidestep the issue by stipulating no true drift

What implications do adaptive designs have for power calculations?

- Power calculations become much more complex, because the central question changes from “What is the probability of declaring a given ATE to be statistically significant at a given alpha level?” to “What is the probability of declaring the best performing treatment arm to be significantly better than the control arm given the allocation and stopping rules?”
- Requires simulation

What implications do adaptive designs have for pre-analysis plans?

- Need to specify what algorithms will be used for allocation and what the stopping rules are (e.g., based on posterior probability thresholds, value remaining, or apparent effectiveness vis-à-vis the control group/standard of care).
- For RCTs in which there may be a change in primary outcome measure after an interim analysis, need to lay out what the criteria will be that will justify a shift in focus. May need a bias correction, too.

Fixed stopping rules

- E.g., stop after a certain number of subjects have been allocated (and use static allocation to a control group)
- Perhaps a better fit for most academic applications
- Still subject to bias! Fortunately, bias diminishes with the number of periods because the best-performing arms gets the most subjects

Empirical applications

- How to word a ballot proposition so as to garner the highest vote share?
- Quite a lot is at stake in terms of how ballot measures are worded, and campaigns seek to select wording strategically
- A wide array of possible wording choices implies a vast multi-arm trial, pitting different wordings against one another

From CNBC.com 22 Jun 2016

The question on the U.K. ballot is stated, "Should the United Kingdom remain a member of the European Union or leave the European Union?" ... U.K. Prime Minister David Cameron accepted a recommendation to change the wording after the phrasing was tested on potential campaigners, academics and language experts.

Test #1: So-called “right to work” ballot measures

- Designed to undermine unions
- Many such ballot measures have been presented to voters in U.S. states over the years, sometimes as amendments to state constitutions
- For purposes of our test, we assembled and standardized the wording from these actual proposals

Example: Right to Work Proposal #4:

Imagine that the following ballot measure were up for a vote in your state. The measure [would amend the State Constitution as follows:] states that no person shall be deprived of life, liberty or property without due process of law. The right of persons to work shall not be denied or abridged on account of membership or nonmembership in any labor union, or labor organization.

If this measure were on the ballot in your state, would you vote in favor or against?

Example: Right to Work Proposal #1:

…prohibit, as a condition of employment, forced membership in a labor organization (union) or forced payments of dues or fees, in full or pro-rata ("fair-share"), to a union. The measure will also make any activity which violates employees' rights provided by the bill illegal and ineffective and allow legal remedies for anyone injured as a result of another person violating or threatening to violate those employees' rights. The measure will not apply to union agreements entered into before the effective date of the measure, unless those agreements are amended or renewed after the effective date of the measure.

Test #2: Minimum wage ballot measures

- Usually declare a minimum wage and sometimes specify indexing for inflation
- Frequently appear on statewide ballots as a tactic to increase voter turnout among left-leaning and low-income citizens
- Again, for purposes of our test, we assembled and standardized the wording from these actual proposals

Example: Minimum Wage Proposal #3:

Imagine that the following ballot measure were up for a vote in your state:

Shall the minimum wage for adults over the age of 18 be raised from current to current + 1 per hour by January 1, 2019?

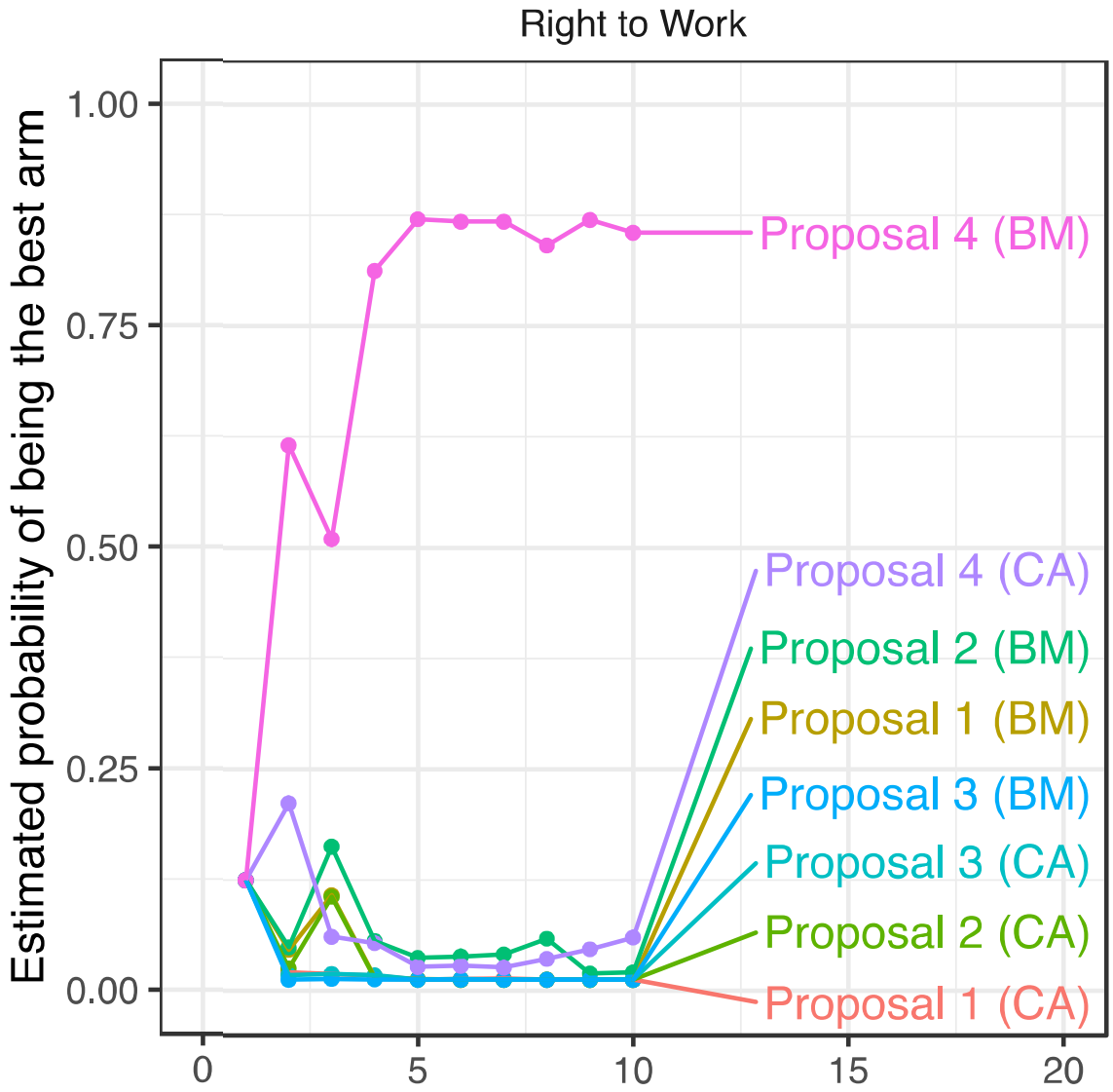
If this measure were on the ballot in your state, would you vote in favor or against?

Example: Minimum Wage Proposal #1:

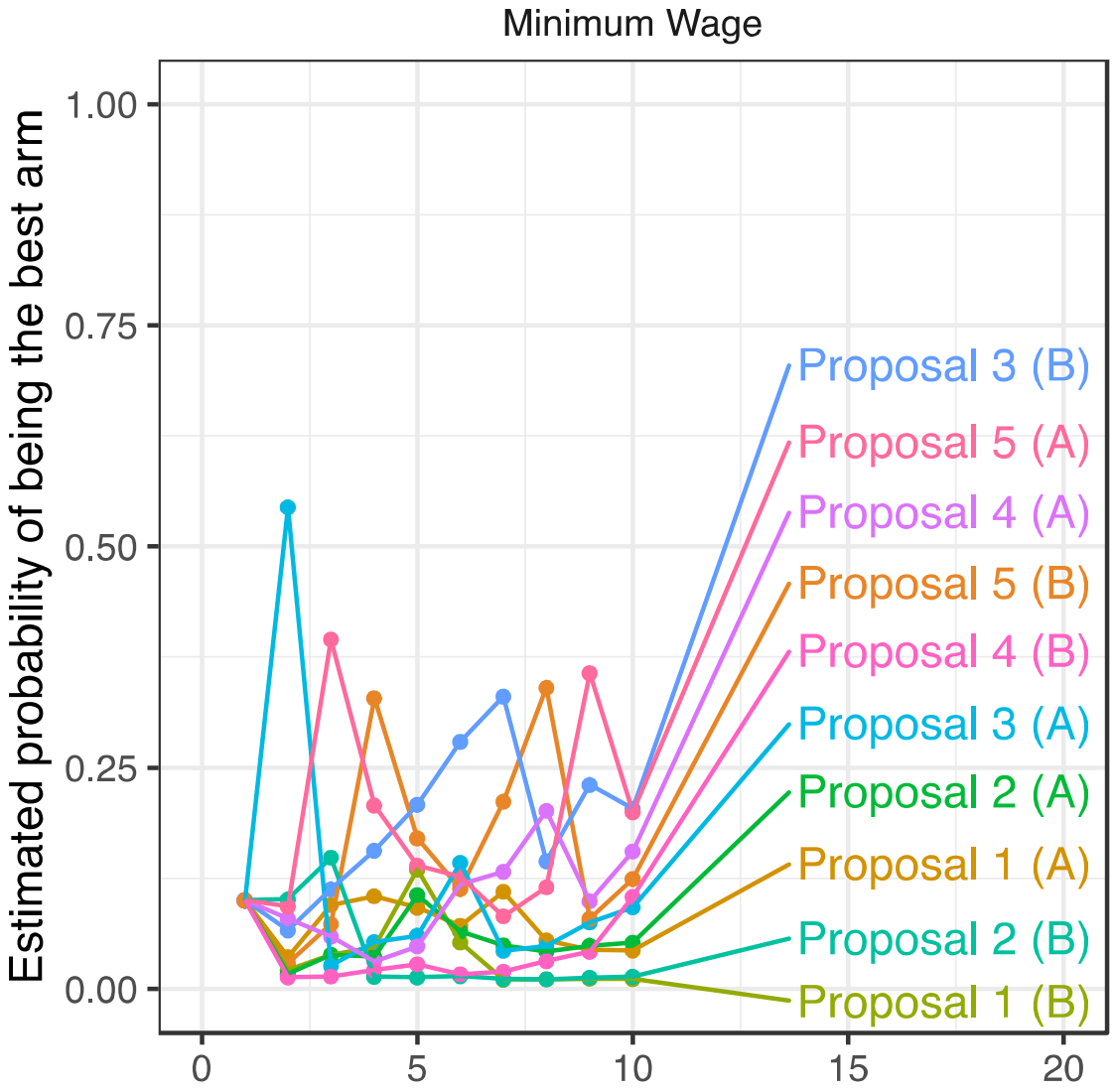
Imagine that the following ballot measure were up for a vote in your state. The measure would increase the minimum wage [from current] to current + 1 per hour, adjusted annually for inflation, and provide that no more than $3.02 per hour in tip income may be used to offset the minimum wage of employees who regularly receive tips.

If this measure were on the ballot in your state, would you vote in favor or against?

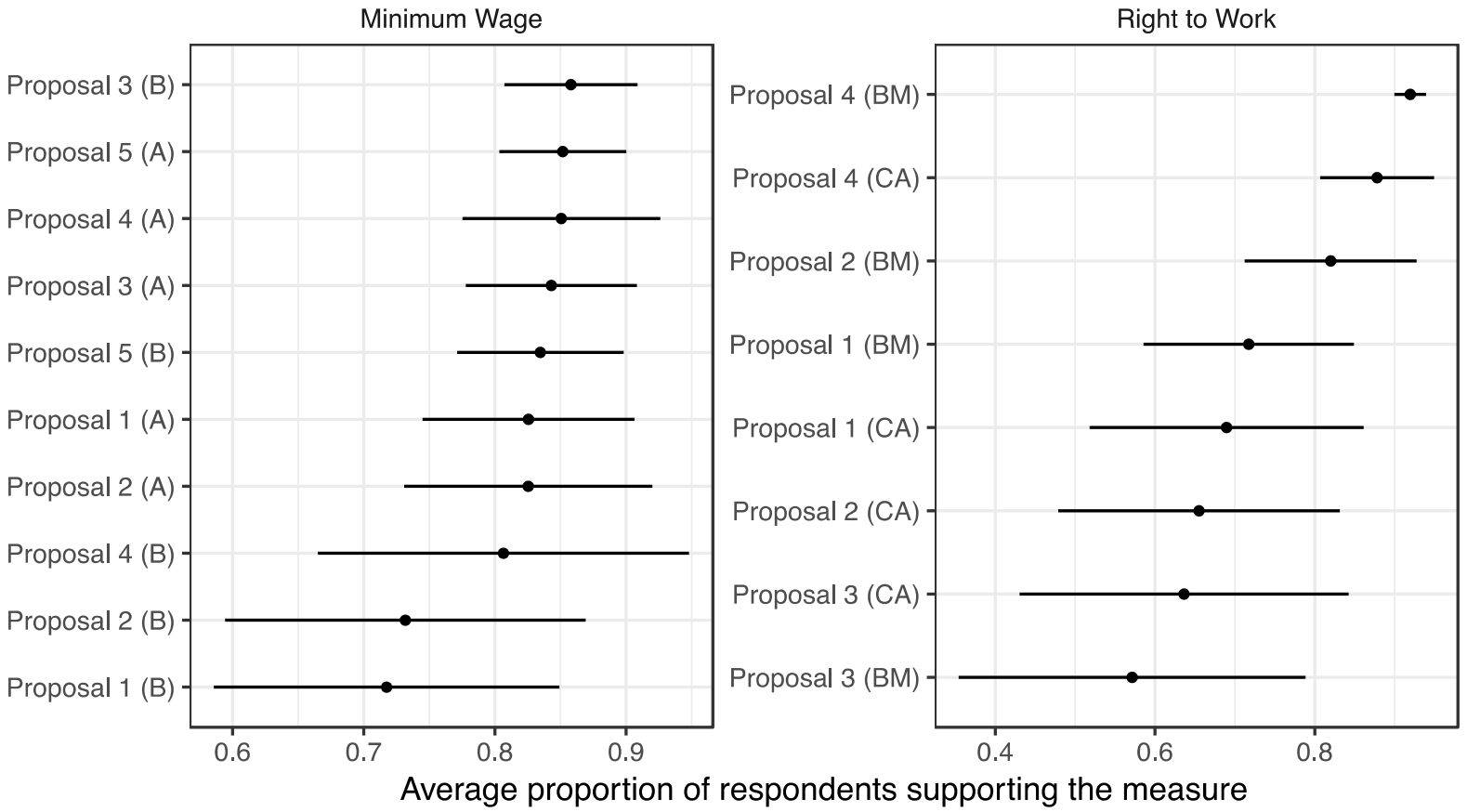
Right to Work Ballot Test: Gains from adaptive design, with SEs that are 45% smaller in the winning arm



Minimum Wage test: No clear winner, with several ballot wordings performing about equally well



Overall popularity of each arm, by type of ballot measure (note confidence intervals widen for less popular arms)



Conclusion

Multi-arm bandits have yet to catch on as a framework for allocating subjects in RCTs

> Villar et al. (2015) Statistical Science: “Despite this apparent near-perfect fit between a real-world problem and a mathematical theory, the MABP has yet to be applied to an actual clinical trial.”

Conclusion

At the same time, there is growing interest in adaptive allocation, spurred by the rapid growth and development of internet sales operations

Many recent technical papers and empirical applications focus on ongoing, high volume internet tests

Despite downside, adaptive design has potential benefits for certain RCTs

- Sifting through large numbers of arms
- Large-scale experiments (e.g., conjoint surveys, correspondence experiments to assess discrimination), where bias is a minor concern
- Designs that are rolled out gradually over time (e.g., in the lab), facilitating out-of-sample validation

Adaptive designs also pose challenges for design and analysis

- Must be carefully regimented ex ante so as to minimize bias and so that the entire procedure can be reproduced via simulation in order conduct proper hypothesis tests

- Not well suited to trials in which one seeks to estimate the effects of all arms, rather than to locate the best arm

- Difficult to implement in the field in small, simultaneous, multi-site trials