Estimating Average Treatment Effects in Block-randomized experiments: A conversation at the No Rules of Thumb Cafe

Jake Bowers — University of Illinois @ Urbana-Champaign
Christopher Grady — University of Illinois @ Urbana-Champaign
Gustavo Diaz — University of Illinois @ Urbana-Champaign

Four field experiments meet at the polished wooden bar of their neighborhood coffee shop.

"Y'all do not look like the bright and energetic agents of evidence-based policy making that I am used to seeing each morning. What's up?" The barista climbs down from dusting the sign hanging over the rows of coffee cups and tea pots lining the wall behind the bar and turns to the customers. The sign says, "No Rules of Thumb."

"We are block randomized today," grumps the first experiment.

"Randomizing treatment assignment within pre-defined strata or blocks was *supposed* to make our findings more precise without adding any more complications to our analyes," sighs the second experiment.

"Yeah. They said, 'A block-randomized experiment is just a series of independent mini-experiments'. And, of course, we know how to estimate treatment effects and standard errors for one experiment, but now we can't agree on how to combine our estimates across more than one block," explains the third.

"And it just doesn't matter how we estimate our effects. We should just always use fixed-effects." grumbles the fourth.

"No! That could be a big problem!" The first experiment almost jumps out of the chair.

"Hey!," the barista blasts a bit of steam from the steam wand and holds up a hand to pause the conflict. "Aren't block randomized experiments just series of small independent experiments? Within each block you should be able to estimate the ATE and SE, right? And you just need to combine those all together somehow, right? So shouldn't it be simple?"

"Ah," the first experiment leans over to look at what the barista is doing, "Right. None of has very small blocks — we each only have three blocks and we have more than 10 people in each block. So we can calculate randomization-justified ATEs and SEs within each block. But we are arguing about the weights. Say I found an effect of 5 in one block and 1 in another block. Should the overall effect just be a simple average that weights each block the same like (5+1)/2=3? What if the first block has 100 people and second block has 10 people? Shouldn't the effect be $(5 \times 100)/110 + (1 \times 10)/110 = 4.64$?"

"So, you are saying that you should combine block-level estimates using the proportion of the total sample in the block as a weight?" the barista is pouring some new coffee beans into bowls set on two different digital scales. “That makes sense to me. Why don’t y’all just do that? You don’t want your overall effect to overrepresent what happened in the small block, right? You have a lot more information about the treatment effect in the bigger block and so that information should play a bigger role in the overall story of the experiment.”

“We should,” says the first experiment a bit grumpily, “After all, lots of previous literature and even fairly simple math tells us that this block-size weighted estimator is unbiased and other weights produce biased estimators. For example, we have been reading Gerber and Green (2012) (Chapter 3) and the blog post, The trouble with ‘controlling for’ blocks or the analytics in Humphreys (2009), which make exactly this argument.”

“By the way, what are you doing with the scales this morning? Usually you just start making our regular drink orders.” The second experiment is also looking at the scales, bowls, and beans.

“Oh. These are new beans. I know that a good espresso depends on striking a balance between bitter and sour. And, in general, with the beans I’m used to getting, I know that about 17 grams of beans ground at about level 2 and extracted at 9 psi for about 24 seconds makes a good double shot using this grinder and this machine and 40 percent humidity in the shope. But, since these are new beans I have to explore the values of these variables a bit before deciding about the right balance. I’ll be back to pulling shots quickly tomorrow. Today, however, I need to make some test coffees that I may not sell in order to learn how to make good coffee with these beans. Just a sec, it will be noisy while I grind this.” With the newly ground beans in the portafilter, the barista puts a scale under a cup, zeros out the weight, touches the button on a timer, and flips the chrome switch on the espresso machine. “So, back to field experiments, why don’t you just use the idea that blocks with more information should contribute more to the estimate and get back to enjoying your day?”

“Well, that is kind of the problem,” says the second experiment a bit wearily, as if rehashing an old argument. “It is very expensive to do policy-relevant field experiments — so we might want to trade a little bit of bias for increases in precision. We also know that a different form of weighting that we call ‘precision-weighting’ is optimal from that perspective.[1]”

“Can you explain the intuition behind the precision-weighting approach?” The third experiment gestures to the barista, “Meanwhile, I’m happy to help you test free espresso shots while you figure out your own procedure for this new batch of beans.”

“Look, take the example of two blocks with treatment effects of 5 and 1 respectively, and imagine that they each have 100 people. But now imagine that the first block had assigned 50 people to treatment and 50 to control, but in the second block the administrators of the program in that block only let you randomize 5 people to treatment. Obviously the first block, with 50 in each of the arms, tells us a lot more about the treatment effect than the second block, with only 5 people in the treatment condition. But both blocks are the same size. And the block-size weights will over-emphasize the effects in the 5 treated-person block from this perspective. From the perspective of information or precision we should instead weight by

[1] See Kalton (1968), page 119 or the exercise on page 128–129 of Gerber and Green (2012), the Green, Coppockm and Lin Lab Standard Operating Procedure or even the derivations for the optimal weight to diminish the size of the variance of a weighted average on wikipedia, or the related derivations in Humphreys (2009) or Hansen and Bowers (2008)).

both block-size **and** proportion assigned to treatment — the second block should get a lower weight in this case. It turns out that these precision weights are exactly this combination. In this case,the first block would get a weight proportional to (100/200)(50/100)(1- 50/100)=.125 — (100/200) is the block-size weight and (50/100) were assigned to treatment. The second block would have a weight proportional to (100/200)(5/100)(1- 5/100)=.0238. After making them sum to 1, we have the precision weights of .84 for the first block and .16 for the second block (compared to .5 and .5 if we only paid attention to block-size). So, we could report the block-size weighted effect of $(5 \times 100)/200 + (1 \times 100)/200 = 5 * (1/2) + 1 * (1/2) = 3$ **or** the precision weighted effect $5(1/2)(50/100)(1 - 50/100) + 1(1/2)(5/100)(1- 5/100) = 5 \times .125 + 1 \times .0238 = 0.65$."

"These are pretty large differences in effect. You said the first one, estimated ATE=3, arises from an unbiased estimator, so we should choose the block-sized weighting in this case, right?" The barista looks puzzled and grinds more beans. The sound of grinding coffee drowns out further conversation for a few seconds. When the grinding stops, the barista asks, "Ok. That makes sense. So, when you have blocks of different sizes or blocks with different probabilities of treatment assignment, you should use that more precise weight, the 'precision-weight,' and when you have blocks where probability of treatment of assignment is the same, both weights should be the same, so then the 'precision-weight' would just be the same as the 'block-size weight.' So you should just use that precision weight always if you want to follow the intuition that overall estimates should reflect blocks with more information, huh? You are actual experiments. I just make coffee. Shouldn't you have figured that out?"

The first experiment shakes their head. "It is not so simple. 'Always use precision weights / fixed effects' might be a rule of thumb that would lead you astray sometimes. And of coure, you should know that there are more than two ways to calculate these weights and that the only way that is guaranteed to produce unbiased estimates is the block-size weight. You should see N. E. Pashley and Miratrix (2020) and N. Pashley and Miratrix (2020) to learn about estimating average treatment effects and their standard errors in all kinds of block randomized experiments — including pair-randomized experiments where you really can't calculate standard errors within pair.[2] We block randomized experiments raise lots of interesting statistical problems." The fourth experiment looks a bit proud saying this. The others chuckle.

"Well. What is the problem then?" The barista tamps the first shot. "You have lots of guidance. Just follow that advice."

"The problem is that we have two rules of thumb to follow given our designs where we have a few large blocks. Some people say that we should 'use fixed effects' (we know that this is the same as saying that we should use precision-weights) and other people say that we should use block-size weights," complains the third experiment.

[2] N. E. Pashley and Miratrix (2020) describes the many different approaches to analyzing experiments with randomization within large blocks (as described in this note) but also within small blocks such as pairs and in sets with only one treated or control unit. The "many small experiments" idea that we describe here is difficult to implement when a block containst only one treated or one control unit since the variance of outcomes for those units is undefined. In a nutshell, they propose a hybrid variance estimator that is a weighted combination of the small- and large-block randomized estimators. For example, fixed effects or precision weighting and block size weighting, but also multilevel models. They also propose an estimator for the variance of block randomized studies when the sizes of blocks differ including blocks with only one treated or one control unit or pairs. N. Pashley and Miratrix (2020) implements the different estimators — allowing analysts to compare their performance in particular cases.

"I see," the barista pulls the first shot and the sound of foaming milk fills the air, "When I was first learning to make espresso and cappuccino, I found a lot of advice on the internet. But, that advice didn't work all the time. I really liked the advice from the chemists — and it gave me good starting places, for example, but nice graphs of acidity by particle size don't really tell me what to do in detail with a given batch of beans. I have to make my own decision about this, knowing that I do not have an ultra clean lab and fancy equipment. So, I've found I sometimes need to play around, try different approaches to see what works. I only do this when I get new beans, or a new grinder, or something else changes, of course. It would be waste of time and money to have to throw away a lot of espresso every morning." The barista pats the gigantic espresso machine with some affection. "Can you do something like this? Try each approach?"

"Hey. This third espresso is better than the first, by the way." The fourth experiment drums fingers on the bar and taps toes on the brass footrest of the bar stool and addresses the other experiments, "I like our barista's idea! We have the DeclareDesign package that we can use. Why don't we compare the two approaches?" The other experiments nod. "We'll take our cappucinos at the table while we work, if you don't mind. Sorry to step away."

"Sure. No problem." The barista motions them away. "I'll be interested to hear what you find."

(TIME PASSES)

"Hey, look at this!" the experiments have come back to the counter carrying their empty coffee cups and holding up one of the their laptops. The morning rush is over. The barista is wiping down the bar.

"Wow! I didn't realize that your designs were so different." The barista stops wiping down the counter. "It is like one of you is a light roast from Ethiopia and the other is a dark roast blend!"

"Yep. It looks like neither rule of thumb worked for all four of us. We even did 10,000 simulations where we re-shuffled treatment assignment within our three blocks differently for each simulation." The first experiment puts down the laptop. "I really should not use the precision-weights or fixed-effects approach. Look at how far off from the truth my estimates would have been! Glad I didn't follow that rule of thumb!"

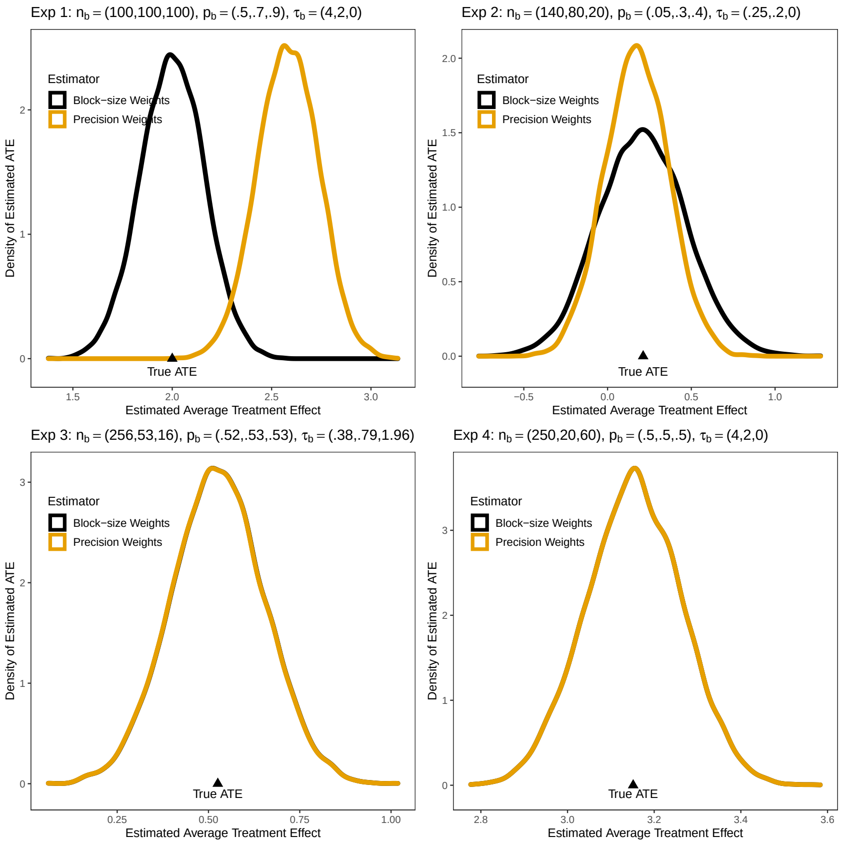
"Is that the gold line top left panel labeled 'Exp 1?'," the barista puts on his reading glasses from the pocket in his work shirt. "You had 100 people in each block, randomly assigned 50%, 70% and 90% of people in each of the blocks to treatment, and the treatment effects were 4, 2 and 0 standard deviations?"

"Yes. My blocks were administrative units that were the same size but one of the adminitrators really wouldn't participate unless nearly everyone got the treatment. The intervention clearly was ultra effective in some places and not in others."

"That gold distribution is so far away from the truth that I wonder whether the confidence intervals created using that as the center would ever cover the truth." The barista leans over the screen.

"Those confidence intervals did a terrible job. In this case precision weights would be terrible – although we also thought this would happen after looking at the simulations using this design from The trouble with 'controlling for' blocks." The first experiment nods.

"But, in my case, precision weights are best," Experiment 2 speaks up. "I also had a situation where the administrators of the blocks differed in their willingness to allow us to randomize but I also had differing block sizes and my treatment effects were nearly the



Exp 1: $n_b = (100,100,100)$, $p_b = (.5,.7,.9)$, $\tau_b = (4,2,0)$

Exp 2: $n_b = (140,80,20)$, $p_b = (.05,.3,.4)$, $\tau_b = (.25,.2,0)$

Exp 3: $n_b = (256,53,16)$, $p_b = (.52,.53,.53)$, $\tau_b = (.38,.79,1.96)$

Exp 4: $n_b = (250,20,60)$, $p_b = (.5,.5,.5)$, $\tau_b = (4,2,0)$

Figure 1: Estimated ATE Distributions for Four Block-randomized experiments. 10,000 simulations of each design holding the population fixed. Precision weights work best for experiment 2, worst for experiment 1, and the same as block-size weights for experiments 3 and 4.

same across the blocks. Here, both confidence intervals have correct coverage, but I can detect a smaller effect if I use the precision weight. This is important to me given my smaller sample size. As you can see, if there is bias, it is tiny and didn't change the coverage of the confidence intervals."

"For both us," the third experiment points to itself and the fourth experiment, "it doesn't matter which weight we use. We were worried for no reason."Both of us were run in administrative units that differed a lot in size, but our administrators were able to randomize more or less the same amount within each block."

"But my treatment effects were just like experiment 1 — one block with no effect and two other blocks with quite large effects. While experiment 3 also had effects that varied, but but the largest effect was in the smallest block, and each block had some effect." The fourth experiment chimed in.

"Well, if you had been in the situations of the other two experiments, you would have had reason to worry. Good thing you did those simulations. And, at least you learned that the block-size weighting approach is unbiased for all four of you — even if it is not as precise as you'd like for experiment 2." The barista takes the empty coffee cups to the sink. "That looked like a lot of work. Would you like another coffee? It would be on the house given how much you've taught me today."

"It was some work, but the simulations were not so difficult to execute. We can put all of the code up on on our github repository so you and your other customers can use it." The fourth experiment opens a computer and starts to type.

The third experiment frowns a litte, "This simulation based approach would work better, by the way, if we used simulation to choose weighting *before* collecting the data. Of course, we can't know in advance what our treatment effects will be, but we could try out some guesses. Most power-analyses specify one effect size for the whole experiment and do not talk about block-by-block effect differences. But maybe we will do more of this if precision gains are really important in the future — like when policy relevant effect sizes are small and we have small samples, or cluster-randomized studies, or relatively light-touch interventions."

"Right," says the first experiment, "So, in our case the conservative thing to do now that we already have collected the data from our experiments would be to use block-size weights even if our confidence intervals could be narrower in experiment 2's case."

The barista smiles. "It is good to know that doing some exploring and simulating works **both** for making a good espresso **and** also for estimating average treatment effects from block-randomized field experiments when you are in a new situation."

References

Gerber, Alan S, and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation.* WW Norton.

Hansen, B. B., and J. Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science* 23: 219.

Humphreys, Macartan. 2009. "Bounds on Least Squares Estimates of Causal Effects in the Presence of Heterogeneous Assignment Probabilities." *Manuscript, Columbia University.*

Kalton, G. 1968. "Standardization: A Technique to Control for Extraneous Variables." *Applied Statistics* 17: 118–36.

Pashley, Nicole E, and Luke W Miratrix. 2020. “Insights on Variance Estimation for Blocked and Matched Pairs Designs.” *Journal of Educational and Behavioral Statistics* XX (X): 1–26. https://doi.org/10.3102/1076998620946272.

Pashley, Nicole, and Luke Miratrix. 2020. *Blkvar: ATE and Treatment Variation Estimation for Blocked and Multisite RCTs.*