Assessing sampling methods for generalization from RCTs: Modeling recruitment and participation

Gleb Furman[1] & James E. Pustejovsky[1]

[1] University of Texas at Austin

Assessing sampling methods for generalization from RCTs: Modeling recruitment and participation

Results

Generalizability

Figure 1 displays the average SMD between the samples and the population for each covariate and at each population participation rate resulting from each sampling method. The dotted horizontal line indicates a cutoff of .25, where SMDs above that indicate large differences between the sample and population for that covariate. Stratified methods consistently performed better than unstratified methods. SBS generally performed as well as or better than SRS. URS often resulted in highly unrepresenative samples except in cases where population participation rates were extremely high.

Figure 2 displays the average B-index for each method across participation rates. At population participation rates of 60% and higher, all methods resulted in similarly generalizable samples. However, at lower rates only SBS and SRS consistantly generated highly generalizable samples. SCS and URS performed equally, while UCS resulted in relatively less generalizable samples.

Feasibility

Figure 3 reports the average number of schools that needed to be contacted before a full sample of $N = 60$ schools was selected. At higher participation rates differences between methods were negligibile. However, as participation rates decreased the disparity between the methods became more apparent. Overall, UCS required the least "effort" to recruit a full sample, followed by URS and SCS, SRS, and finally SBS. Figure 4 plots the participation rates of schools approached for recruitment against the population participation rates. As expected, URS participation rates reflected those in the population. Both UCS and SCS resulted in higher participation rates, while SRS and SBS resulted in lower participation

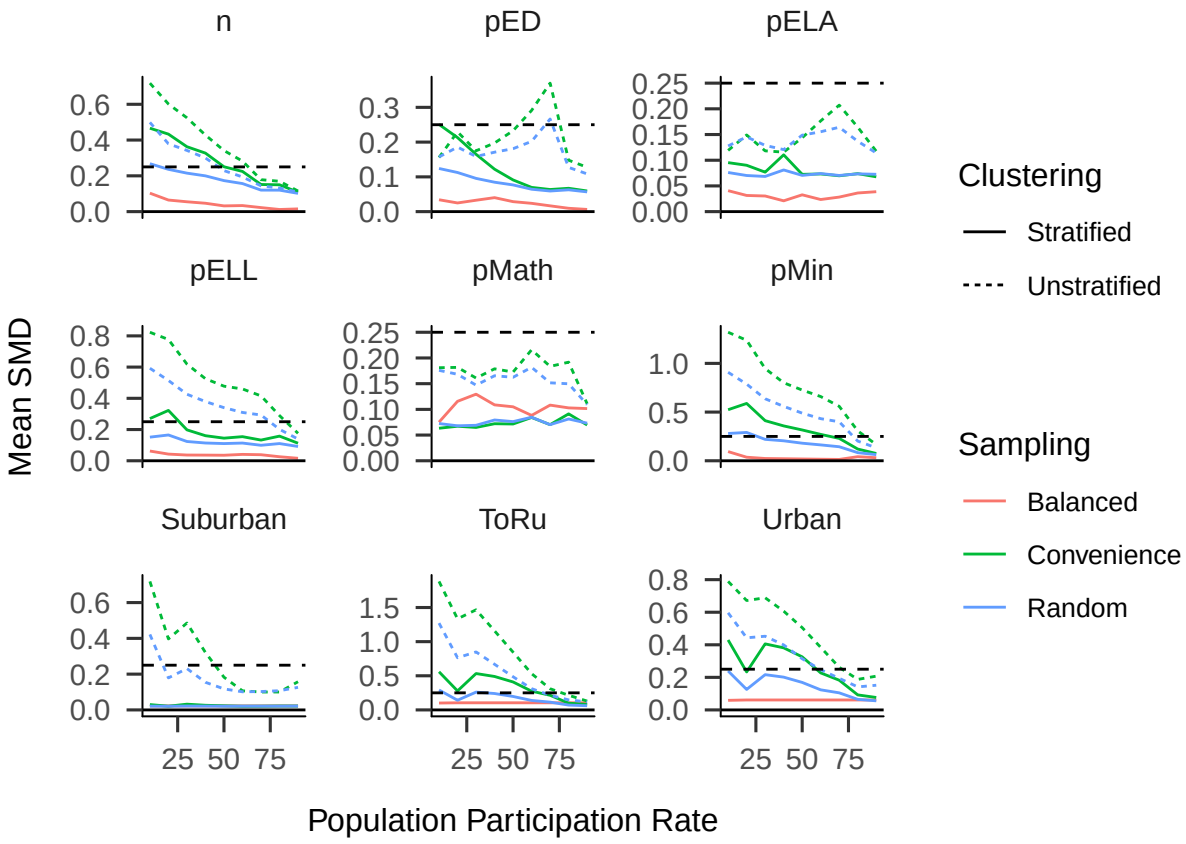

Figure 1. Averge Standardized Mean Differences between sample and population rates.

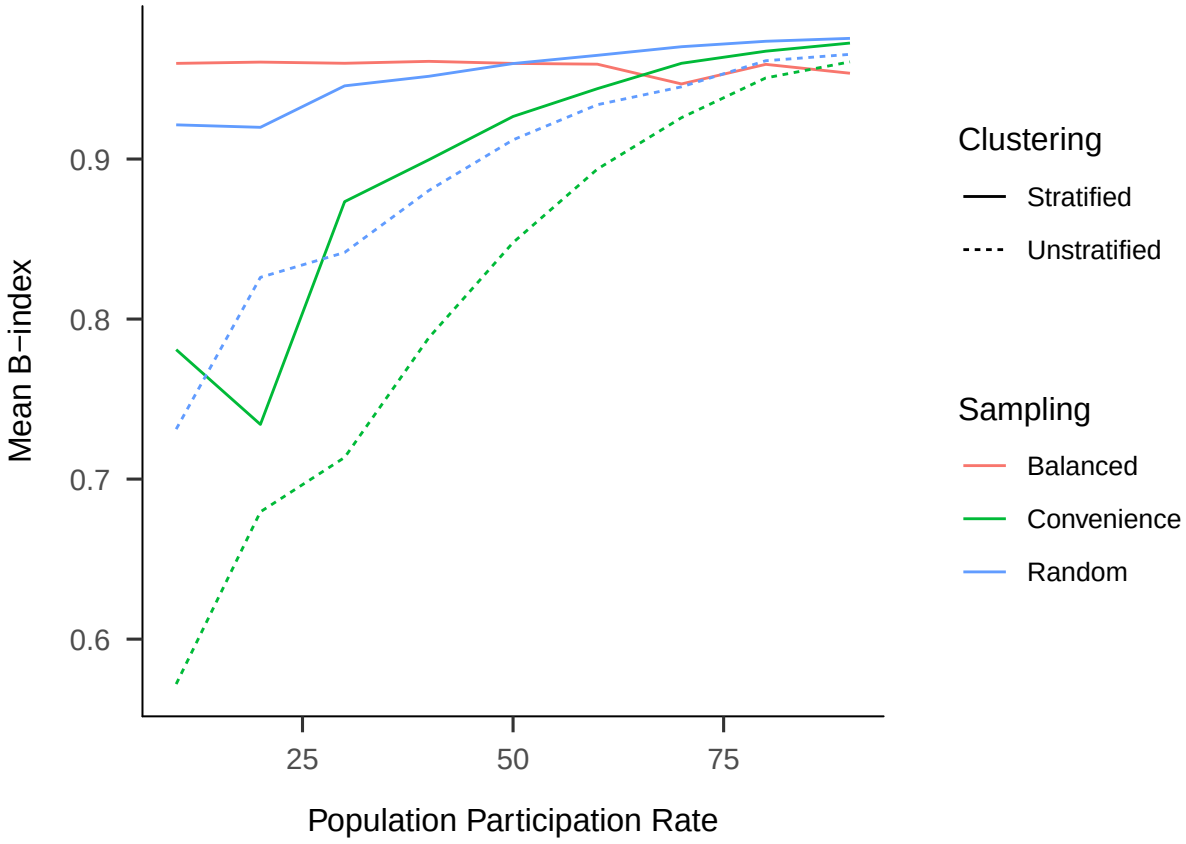

Figure 2. Averge B-index across participation rates by sampling method

Discussion

The main goal of this study was to lay a groundwork for exploring the effectiveness and feasibility of sampling methods in the educational context. Our work uncovered several limitations in designing sampling and participation models. The sampling methods we developed made several assumptions about researcher and recruiter behavior in selecting a sample. First, that recruiters always prioritize schools that are most likely to participate. In truth, many other factors play a role such as proximity of sample units to the researcher and to each other, existing relationships between the recruiters and the sample units, and other researcher assumptions about the sample unit's characteristics. Second, that recruiters have approximiate knowledge of how likely a sampled unit is to participate. Though researchers may speculate about units that are more willing to participate (schools in larger urban

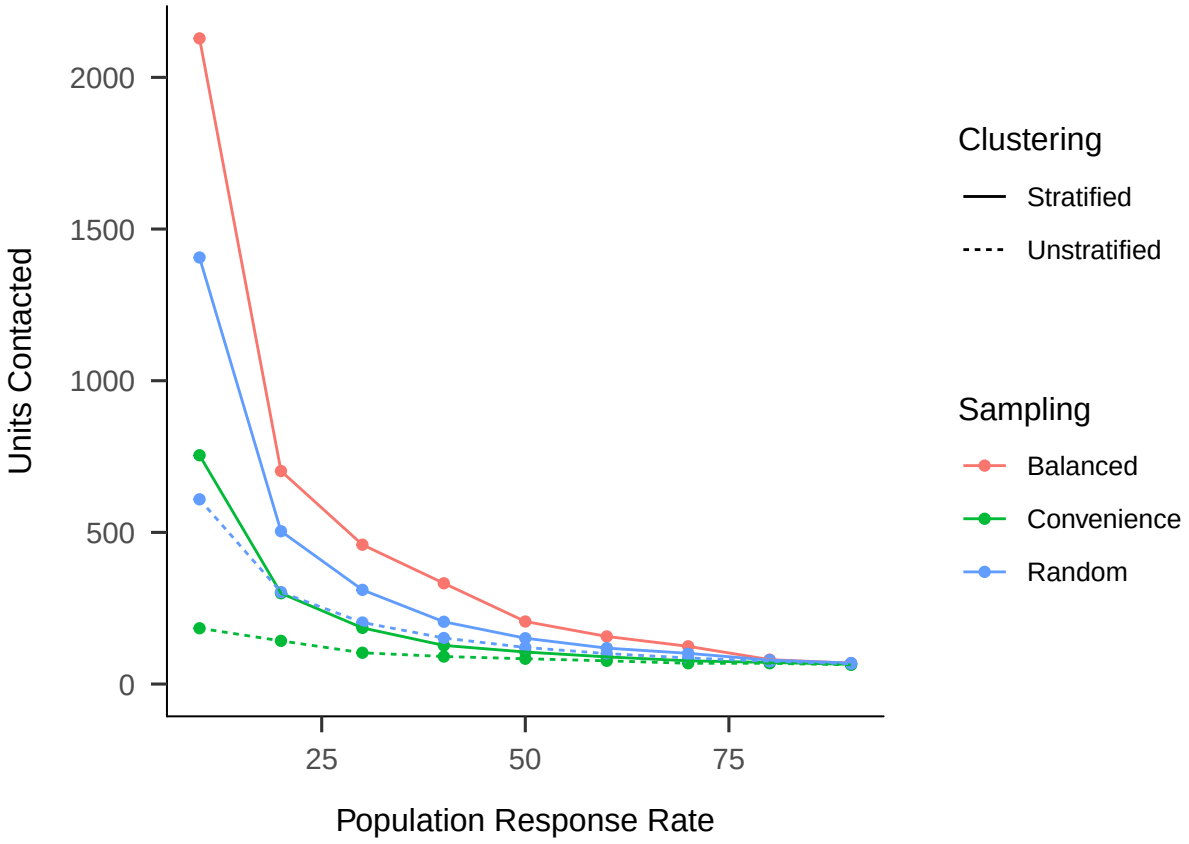

Figure 3. Averge number of schools contacted to achieve N = 60

distrits) and prioritize recruiting such units, it is not likely that they would estimate willingness as well as we have simulated. Given this, it is possible that the "feasibility" of the convenience methods is overestimated.

That said, our participation model is also likely innacurate. The parameters in our response generating model are based on values from a study that examined the difference between schools participating in large-scale randomized control trials (RCT) and the overall population of schools. However, these RCTs themselves typically rely on some form of convenience sampling. In that sense our parameters reflect participation rates of schools that are likely to participate in RCTs, rather than the full population of schools. Furthermore, the decision of whether a school participates in such a study is multi-leveled. Generally, districts serve as gatekeepers, requiring research requests to be submitted and approved

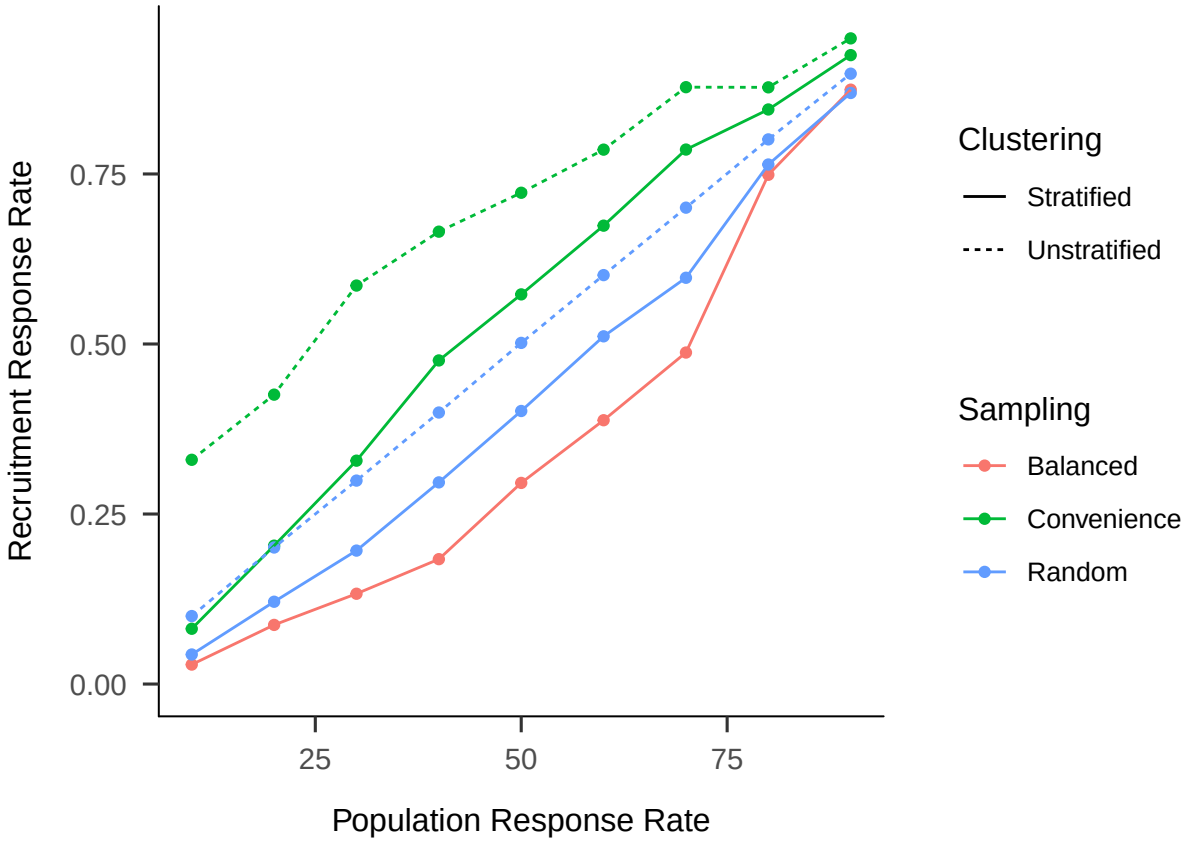

Figure 4. Recruitment response rates for each sampling method

before recruitment can begin. If the request is denied, no schools within that district may be recruited. If approved, schools may be contacted individually. The ultimate decision may then rest with administrators, or they may be passed on to teachers most affected by the study.

Despite these limitations, we believe that our findings reasonably represent the relative performance of the various sampling methods we tested in the context of educational research. In terms of selecting a generalizable sample, SBS results in a considerable improvement compared to UCS. However, given the difficulty with which those samples are recruited, SBS is unlikely to be fully implemented in the ideal form. Instead, SCS may be a reasonable compromise. Our findings indicate that any sampling method is greatly improved by first stratifying the population. We show that performing a convenience sample within

strata (SCS) is comparable to simple random sampling (URS) both in terms of generalizability and feasibility. URS is considered the gold standard for generalizabilty, however it is not typically implemented due to it requiring substantial resources and other concerns of practicality. The advantage of SCS here is that it enables researchers to mitigate recruitment costs by targeting schools from the same strata that are in close proximity to each other. Beyond generalizability, stratifying in this manner also offers the additional advantage of transparency by forcing the researcher to make sampling decisions in the study design phase, and to keep track of sampling decisions as recruitment is implemented.

Finally, we hope to make clear the disadvantages of convenience sampling in this context. Large scale MRTs are expensive to implement, however by not investing in robust recruitment strategies researchers severly limit the impact and relevance of of their work. We believe that the increased cost of a sampling method designed for generalizability is greatly outweighed by the benefit of an intervention whose impacts we can estimate more accurately and for a wider population.

References