Abstract Title Page

Title: Site selection in experiments: A follow-up evaluation of site recruitment in two scale-up studies

Authors and Affiliations:
Elizabeth Tipton, *Teachers College, Columbia University*
Lauren Fellers, *Teachers College, Columbia University*
Sarah Caverly, *SEDL*
Michael Vaden-Kiernan, *SEDL*
Geoffrey Borman, *University of Wisconsin, Madison (gborman@education.wisc.edu)*
Kate Sullivan, *SEDL (kate.sullivan@sedl.org)*
Veronica Ruiz de Castillo, *SEDL (veronica.ruizdecastilla@sedl.org)*

Abstract Body

Background / Context:

Randomized experiments are commonly used to evaluate if particular interventions improve student achievement. In these evaluations, the goal is typically to estimate a single, average treatment impact, and ideally, the results of the evaluation can be used to make policy decision by schools, districts, and other governing bodies (e.g., via the What Works Clearinghouse). While random assignment to the treatment conditions ensures that that the treatment in fact *causes* these changes, typically the schools or districts that take part in the experiment are not randomly selected from a well-defined inference population. If an intervention is more or less effective in some schools or districts than others, however, this convenience sampling strategy results in a causal effect that does not readily generalize.

Recently, attention has turned to developing new methodologies for improving generalizations from large-scale experiments (see Schochet, Puma, & Deke, 2014). There have been three streams of research in this area. The first has focused on *assessing* the degree of similarity between the convenience sample of schools or districts in a completed experiment (e.g., Stuart, Cole, Bradshaw, & Leaf, 2011; Olsen, Orr, Bell, & Stuart, 2013; Tipton, in press). The second area focuses on *reweighting* this convenience sample to be more similar to one or more well-defined inference populations (e.g., O'Muircheartaigh & Hedges, 2014; Tipton, 2013). This work shows that there can often be a large-penalty to delaying discussions of generalization until after the evaluation is complete – increased standard errors and, often, limits to bias reduction. In reaction to these limitations, the third area shifts focus from improvements through statistical adjustments to improvements through design and improved recruitment strategies (e.g. Tipton et al, 2014; Tipton, 2014; Roschelle et al, 2014).

Tipton et al (2014) provide a purposive sampling alternative to the convenience sampling most commonly found in the field. This design-based approach uses propensity score methodology to first compare an inference population to those eligible for recruitment in the experiment, and then creates strata for site-selection[*]. The goal is to help recruiters create a recruitment strategy that is targeted and, that when perfectly implemented, results in a sample of sites that is like a miniature of the inference population of interest. When not perfectly implemented (which is seen as likely), the goal is to reduce or eliminate the under-coverage problems that limit the effectiveness of post-hoc statistical adjustments. The paper situates this more general method in relation to two scale-up studies conducted by SEDL and the University of Wisconsin, Madison: one of Open Court Reading and the other of Everyday Math. These studies began recruitment in the fall of 2011, with the first round of experimental results available in the spring of 2014.

Purpose / Objective / Research Question / Focus of Study:

This paper is a follow up study to the examples proposed and carried out in Tipton et al (2014), with the goal of evaluating the success of these methods in practice, as well as addressing additional problems that arose in recruitment. The three aims of this work are: 1) Comparing sites actually included in the final study sample to those sites that were proposed in the original

[*] Note that this work assumes that not all units in the population are eligible to be in the experiment. When this is not the case, Tipton (2014) provides an alternative stratification method using cluster analysis.

stratification plans; 2) Discussing issues of "non-response" that arose in recruitment, whereby districts were contacted but declined to take part in the study and wherein the list of eligible districts needed to be replenished; and, 3) Utilizing re-weighting methods (Tipton, 2013) to calculate final estimates of the population average treatment effect for the originally intended population. By looking at these methods as they are implemented in real time experiments we are able to discern issues with sample selection, site recruitment, and problems within the analysis plan.

Significance / Novelty of study:
This work is important for three reasons. First, Tipton et al was the first study to implement a propensity score based stratified selection plan, and this paper extends that work. Second, by providing information on the proposed recruitment plan, as well as issues that arose in recruitment, we intend to provide feedback to the field regarding the real constraints and issues impeding generalization. Third, this work also provides an opportunity to explore the issues of non-response typically studied in survey sampling to the problem of site-recruitment in experiments.

Statistical, Measurement, or Econometric Model:
Background

Previously in Tipton et al (2014) a new multistep method for site selection was proposed. This method calls for: 1) Defining an inference population; 2) Defining requirements of eligibility; 3) Selecting covariates that might be related to the variability of treatment effects; 4) Creating strata and allocating the sample proportionally to those strata; and finally, 5) Recruiting sites with a goal of including at least some of the eligible sample in each stratum. The recruitment plans created for the OCR and EM studies originally proposed to include 15 districts, each with four schools, where treatment schools serve as control groups for the opposite program. This study focused on elementary schools only. Eligibility for this study was defined as those schools that had not used the program in the previous three years and were not missing data.

In the OCR and EM studies, inference populations were defined as current users of these curricula. Data was gathered from the developer (McGraw Hill) for these populations. An interesting problem was that it was impossible to include any units in the population in the actual study (because they were already users of the program). The goal was then to select a sample of sites who were most "like", i.e., compositionally similar, to those in the defined population. Twelve covariates were selected from the Common Core of Data (seen in **Table 1**). Both the OCR and EM populations were divided into three strata each. Because one sample was needed to be selected for both evaluations (experiments were combined), these three by three strata were combined into nine total strata, and eligible sites were identified within each. Eligible sites within each stratum were ranked in terms of similarity to the stratum means. More information is available in Tipton et al (2014).

Comparison on final sample to sampling plan

The beginning of this paper compares those schools in the final selection of sites in both scale-up experiments to original strata proposed in Tipton et al. Researchers experienced several difficulties in recruitment that impacted site selection for some strata. First, some districts were able to provide more schools than were needed, while some districts were only interested in

being included in *either* the OCR or EM experiment, but not both. Overall there were nine districts across both studies: two districts were in the OCR study only, two in the EM study only, and five were in both. Of these nine districts, only six were originally eligible to be included in the experiment. Of those three that were ineligible, two were originally excluded because of missing Common Core Data covariates, and the third was originally excluded because of previous program use (because recruitment happened in multiple stages, changing later eligibility based on program use).

In **Figure 1**, we include the original graphic from Tipton et al, identifying eligible districts in the nine strata, and now highlighting the districts involved in the final study. Here the grey dots represent eligible schools as they fall into each of the nine strata. The larger shapes (see figure key for distinction between OCR and EM experiment schools) are those nine districts included in the final study. Solid black lines are used to show the divisions of each stratum. As seen in this figure sites are represented in all three marginal strata for the OCR study, however for EM sites one marginal stratum does not have an eligible site included in the final study. For those three districts not originally eligible, missing data for two districts was obtained or imputed and used in the original logistic regression model to predict their stratum membership. Strata lines are drawn to allow for equal population size within each stratum, meaning that marginal distributions each contain 1/3 of the associated population. For the EM experiment a dashed line is included to represent if the stratum line was moved so that each stratum contains at least one district. Doing so shifts the population proportion so that the first stratum now contains 23.09%, the second contains 43.55%, and the third contains 33.33%.

Evaluation of the final sample: Balance

The purpose of the strata is to lead to a balanced sample where the sampled districts are compositionally similar to the inference population of districts. Therefore it is important to evaluate this for each covariate originally selected. In **Table 1**, we compare the population to the sample for OCR. In **Table 2**, we compare population and sample means for EM. In all comparisons, we use the mean difference, standardized by the population standard deviation. The last column in each of the previous tables included the effect of post-hoc reweighting on measures of balance. We utilize the methods proposed in Tipton (2013) for this procedure. The mean estimator for each covariate is calculated as,

$$\bar{X}_{sub} = \sum_{j=1}^{3} w_{p_j} X_j \ ,$$

where X_j is the district average value of X for those in stratum j. Typically reweighting reduces bias in the covariates.

From these tables we can see the population and sample means for each of the selected covariates (some of the 12 selected covariates are measured separately i.e., district ethnicity). In the OCR experiment we see that we increase balance on 12 out of 21 variables. Balance is improved on all variables regarding district geography and urbanicty, as well as most variables regarding student ethnicity. All but one variable in the selected sample have means less than one standard deviation from the population mean. For EM we see 10 that achieved better balance however this could be due to the recruitment of non-eligible schools. For those variables regarding student attributes (ethnicity, ELL and F/R lunch status) balance was improved on most. District level variables saw balance remain the same or improve after reweighting.

"Non-response" and recruitment issues

In this paper we also include a discussion of those sites who agreed to be in the experiment as compared to those that did not wish to participate. The concern with non-response is one that has yet to be addressed in current literature. **Table 3 and Table 4** (OCR and EM populations are evaluated separately) compare those eligible schools and the defined inference populations. Within these tables we see that there are different means for each of the recruitment decisions (Not recruited for the study; Did not respond to recruitment communication; Did not want to participate in the study; and Agreed to be in the study). From the steps enumerated previously, a detailed list of 675 eligible schools was created and recruitment was aimed at these schools. Due to low numbers of schools who agreed to be in the study, recruitment happened in three stages. Only the first wave of recruitment is addressed here, later stages will be assessed in the full paper. From these tables we see that for most variables (categories of urbanicity showed larger differences between the population and those who refused to be in the study), schools that were not recruited had the largest deviation from the population mean. Those who refused to be in the study also show the largest between-group mean differences from those who were not recruited or did not respond to recruitment communication.

In the full paper, we apply methods of non-response analysis adopted from survey sampling to the problem. Results of this are not addressed here for issues of space.

Reweighted estimators of the ATE

Finally, in the paper we will include reweighted estimates of treatment effects. We do not yet have final outcomes, but expect to receive them within the next few weeks. Since we have site membership for strata, we can easily estimate these results at a later date, using the post-stratification estimator,

$$T_{sub} = \Sigma \underline{w}_{pj} T_j,$$

where T_j is the district average ATE in stratum j, and w_{pj} is the population weight. Note that in the OCR study, these weights are equal ($w_{pj} = 1/3$), whereas in the EM population, these weights differ (see Figure 1).

Usefulness / Applicability of Method:

Throughout, we situate the paper in relation to the Open Court Reading and Everyday Math scale-up evaluations conducted by SRI. By focusing on the specific problems encountered in site recruitment, we offer both feedback on site-selection methods recently developed and propose additional methods for improving generalizations.

Conclusions:

This paper follows up on previously proposed methods for stratified sampling based on covariates that could explain variability in treatment effects. It provides a real time example of how these methods can be used and how they enable researchers to generalize. This paper evaluates the implementation of these plans and their success for achieving better balance on selected covariates. Within this paper we also highlight areas where problems with recruitment or methodology occurs and how it can be better addressed in future implementations of these procedures.

Appendices

Appendix A. References

Olsen, R.B., Orr, L.L., Bell, S.H., and Stuart, E.A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management, 32* 107-121.

O'Muircheartaigh, C., Hedges L.V. (2014). Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society, Series C* (Applied Statistics), 63: 195-210.

Roschelle, J., Feng, M., Gallagher, H.A., Murphy, R., Harris, C., Kamdar, D., & Trinidad, G. (2014). Recruiting participants for large-scale random assignment experiments in school settings. Menlo Park, CA: SRI International.

Schochet, P. Z., Puma, M., & Deke, J. (2014). Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods (NCEE 2014–4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from http://ies.ed.gov/ncee/edlabs.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A*, Part 2, 369-386.

Tipton, E. (*in press*). How generalizable is your experiment? Comparing a sample and population through a generalizability index. *Journal of Educational and Behavioral Statistics.*

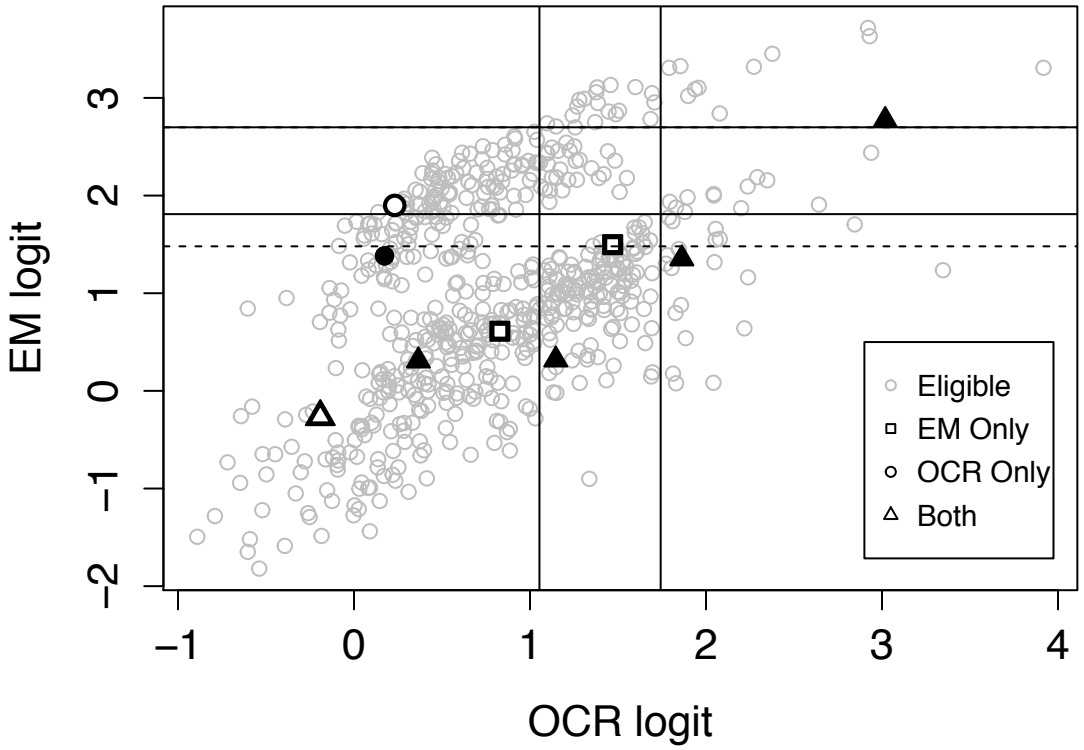
Tipton, E. (2014). Stratified sampling using cluster analysis: A balanced-sampling strategy for improved generalizations from experiments. *Evaluation Review, 37*(2): 109-139.

Tipton, E. (2013). Improving generalizations from experiments using subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics, 38*: 239-266.

Tipton, E., Hedges, L.V., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiements: A New method using propensity score stratified sampling. *Journa of Research on Educational Effectiveness 7(1)*, 114-135.

Appendix B. Tables and Figures
Not included in page count.

Figure 1: Districts in OCR and EM Studies



Notes: Solid lines denote strata originally proposed for study recruitment. Dashed lines indicated re-drawn strata for reweighting procedure.

Table 1: Comparison of OCR Population and Sample Means

		OCR Population		OCR Experiment			
		N=1902		n=7		Pre	Post
Category	Covariate	M	SD	M	SD	SMD	RWT SMD
	Average number of	9180.988	27507.737	9956.000	3418.000	0.028	0.017
Student	students in the district						
	Race/ethnicity of district						
	% White	0.584	0.308	0.684	0.082	0.324	0.115
	% Hispanic	0.200	0.251	0.166	0.069	0.133	0.418
	% Black/African American	0.108	0.181	0.064	0.003	0.243	0.427
	% other	0.108	0.151	0.086	0.010	0.149	0.419
	% students ELL	0.086	0.130	0.134	0.049	0.368	0.306
	% students F/RL	0.433	0.226	0.465	0.040	0.143	0.798
District	Urbanicity of districts						
	% Urban	0.093	0.291	0.500	0.500	1.401	0.253
	% Suburban	0.219	0.414	0.500	0.500	0.679	0.128
	% Town or Rural	0.688	0.463	0.000	0.000	1.484	0.045
	Geographic location						
	% Northeast	0.150	0.357	0.000	0.000	0.421	0.421
	% Midwest	0.175	0.380	1.000	0.000	2.175	0.021
	% South	0.361	0.480	0.000	0.000	0.752	0.636
	% West	0.314	0.464	0.000	0.000	0.676	0.317
	District Revenue	12624.068	6008.853	10418.311	720.391	0.367	0.621
Community	Educational Attainment						
	% Grade 8 or lower	0.116	0.097	0.068	0.032	0.494	0.008
	% <HS grad	0.161	0.062	0.128	0.030	0.540	0.673
	% HS grad	0.387	0.105	0.410	0.021	0.221	0.390
	% Postsecondary	0.336	0.172	0.394	0.082	0.338	0.477
	% 5-17 year olds in poverty	0.165	0.110	0.098	0.051	0.604	0.445
Census area	% labor force	0.622	0.078	0.706	0.024	1.079	0.907
financials	Median income (overall)	48538.326	19108.375	52570.000	6822.000	0.211	0.582

Notes: ELL - English Language Learner status; F/RL - Free or Reduced Price Lunch status

Table 2: Comparison of EM Population and Sample Means

		EM Population		EM Experiment			
		N=3118		n=7		Pre	Post
Category	Covariate	M	SD	M	SD	SMD	SMD
Student	Average number of students in the district	6790.280	22350.944	11851.429	10059.213	0.226	0.063
	Race/ethnicity of district						
	% White	0.741	0.256	0.487	0.268	0.996	0.836
	% Hispanic	0.091	0.145	0.080	0.087	0.077	0.016
	% Black/African American	0.087	0.162	0.323	0.325	1.452	0.560
	% other	0.080	0.118	0.111	0.102	0.257	1.024
	% students ELL	0.043	0.080	0.024	0.021	0.236	0.003
	% students F/RL	0.355	0.213	0.554	0.259	0.936	1.323
District	Urbanicity of districts						
	% Urban	0.077	0.267	0.143	0.350	0.247	0.144
	% Suburban	0.306	0.461	0.143	0.350	0.354	0.413
	% Town or Rural	0.617	0.486	0.714	0.452	0.200	0.313
	Geographic location						
	% Northeast	0.248	0.432	0.000	0.000	0.574	0.574
	% Midwest	0.441	0.497	0.000	0.000	0.888	0.656
	% South	0.155	0.362	0.714	0.452	1.544	1.553
	% West	0.156	0.363	0.286	0.452	0.357	0.029
	District Revenue	13037.115	5403.676	8798.514	3981.296	0.784	0.758
Community	Educational Attainment						
	% Grade 8 or lower	0.077	0.059	0.110	0.073	0.555	0.603
	% <HS grad	0.135	0.060	0.182	0.064	0.788	1.273
	% HS grad	0.410	0.119	0.378	0.116	0.265	0.121
	% Postsecondary	0.378	0.184	0.330	0.216	0.262	0.682
	% 5-17 year olds in poverty	0.117	0.090	0.222	0.096	1.158	1.136
Census area financials	% labor force	0.650	0.071	0.564	0.073	1.219	1.520
	Median income (overall)	54747.237	21059.787	43312.143	19428.440	0.543	0.882

Notes: ELL - English Language Learner status; F/RL - Free or Reduced Price Lunch status

Table 4: OCR Recruitment Means compared to Population means

		Not Eligible	Eligbile for EM and OCR Study							
		Population	Not Recruited		No Response		Said NO		Said YES	
		N=1902	n=95		n=369		n=152		n=3	
Category	Covariate	M	M	SMD	M	SMD	M	SMD	M	SMD
Student	Average number of students in the district	9180.988	13123.126	0.143	11379.539	0.080	10205.401	0.037	13303.667	0.150
	Race/ethnicity of district									
	% White	0.584	0.495	0.290	0.527	0.186	0.601	0.054	0.665	0.262
	% Hispanic	0.200	0.252	0.207	0.231	0.127	0.188	0.047	0.077	0.486
	% Black/African American	0.108	0.144	0.202	0.138	0.167	0.116	0.043	0.165	0.314
	% other	0.108	0.109	0.006	0.103	0.032	0.096	0.083	0.093	0.103
	% students ELL	0.086	0.124	0.293	0.101	0.114	0.085	0.006	0.026	0.465
	% students F/RL	0.433	0.462	0.131	0.436	0.015	0.384	0.215	0.571	0.614
District	Urbanicity of districts									
	% Urban	0.093	0.221	0.441	0.225	0.454	0.151	0.201	0.333	0.827
	% Suburban	0.219	0.326	0.259	0.350	0.315	0.441	0.535	0.000	0.530
	% Town or Rural	0.688	0.453	0.507	0.425	0.566	0.408	0.604	0.667	0.045
	Geographic location									
	% Northeast	0.150	0.095	0.156	0.176	0.072	0.237	0.242	0.000	0.421
	% Midwest	0.175	0.179	0.012	0.138	0.096	0.322	0.389	0.000	0.460
	% South	0.361	0.326	0.073	0.355	0.013	0.184	0.368	0.667	0.636
	% West	0.314	0.400	0.186	0.331	0.036	0.257	0.123	0.333	0.042
	District Revenue	12624.068	10841.593	0.297	11210.030	0.235	12221.975	0.067	10532.010	0.348
Community	Educational Attainment									
	% Grade 8 or lower	0.116	0.106	0.100	0.104	0.122	0.084	0.330	0.140	0.253
	% <HS grad	0.161	0.159	0.035	0.159	0.027	0.139	0.351	0.206	0.719
	% HS grad	0.387	0.361	0.250	0.365	0.211	0.374	0.120	0.457	0.659
	% Postsecondary	0.336	0.374	0.222	0.372	0.208	0.403	0.385	0.198	0.805
	% 5-17 year olds in poverty	0.165	0.144	0.191	0.141	0.213	0.116	0.442	0.224	0.544
Census area financials	% labor force	0.622	0.645	0.291	0.644	0.277	0.652	0.382	0.510	1.444
	Median income (overall)	48538.326	53007.400	0.234	52913.702	0.229	58282.836	0.510	35735.333	0.670

Notes: Groups are separated by their recruitment decision; population (not eligible for recruitment), not recruited, no response, said no, or said yes.

Table 4: EM Recruitment Means compared to Population means

		Not Eligible	Eligbile for EM and OCR Study							
		Population	Not Recruited		No Response		Said NO		Said YES	
		N=3118	n=95		n=369		n=152		n=3	
Category	Covariate	M	M	SMD	M	SMD	M	SMD	M	SMD
Student	Average number of students in the district	6790.280	13123.126	0.283	11379.539	0.205	10205.401	0.153	13303.667	0.291
	Race/ethnicity of district									
	% White	0.741	0.495	0.963	0.527	0.837	0.601	-0.549	0.665	0.298
	% Hispanic	0.091	0.252	1.106	0.231	0.967	0.188	0.666	0.077	0.094
	% Black/African American	0.087	0.144	0.352	0.138	0.313	0.116	0.175	0.165	0.477
	% other	0.080	0.109	0.245	0.103	0.196	0.096	0.131	0.093	0.105
	% students ELL	0.043	0.124	1.020	0.101	0.728	0.085	0.533	0.026	0.213
	% students F/RL	0.355	0.462	0.504	0.436	0.382	0.384	0.137	0.571	1.018
District	Urbanicity of districts									
	% Urban	0.077	0.221	0.541	0.225	0.555	0.151	0.279	0.333	0.962
	% Suburban	0.306	0.326	0.044	0.350	0.095	0.441	0.293	0.000	0.664
	% Town or Rural	0.617	0.453	0.338	0.425	0.394	0.408	0.430	0.667	0.102
	Geographic location									
	% Northeast	0.248	0.095	0.354	0.176	0.166	0.237	0.025	0.000	0.574
	% Midwest	0.441	0.179	0.528	0.138	0.610	0.322	0.239	0.000	0.888
	% South	0.155	0.326	0.472	0.355	0.552	0.184	0.080	0.667	1.412
	% West	0.156	0.400	0.672	0.331	0.480	0.257	0.277	0.333	0.488
	District Revenue	13037.115	10841.593	0.406	11210.030	0.338	12221.975	0.151	10532.010	0.464
Community	Educational Attainment									
	% Grade 8 or lower	0.077	0.106	0.483	0.104	0.447	0.084	0.106	0.140	1.061
	% <HS grad	0.135	0.159	0.405	0.159	0.414	0.139	0.077	0.206	1.191
	% HS grad	0.410	0.361	0.413	0.365	0.379	0.374	0.297	0.457	0.394
	% Postsecondary	0.378	0.374	0.020	0.372	0.033	0.403	0.133	0.198	0.979
	% 5-17 year olds in poverty	0.117	0.144	0.295	0.141	0.268	0.116	0.011	0.224	1.189
Census area financials	% labor force	0.650	0.645	0.075	0.644	0.092	0.652	0.024	0.510	1.973
	Median income (overall)	54747.237	53007.400	0.083	52913.702	0.087	58282.836	0.168	35735.333	0.903

Notes: Groups are separated by their recruitment decision; population (not eligible for recruitment), not recruited, no response, said no, or said yes.