Michigan Teen Smoking and Drug Use Survey Sample Design

Kevin Linares
and
Jianing Zou
and
Weishan Jiang
and
Xiaoqing Liu
University of Maryland

April 22, 2025

Abstract

1 Introduction

The State of Michigan Department of Education (MDE) requires data to monitor teenage smoking and drug use, partly to assess compliance with tobacco industry settlements. This report details the design of a statewide probability sample of Michigan teenagers (enrolled students in grades 7-12) developed to meet the Department's needs. The objective was to create a cost-effective sample design capable of producing estimates for key variables with desired mean and sampling variance for both the state overall and each of the nine official school regions. Based on cost considerations and a review of alternatives, a two-stage school-based sample design was chosen by the client. We outline in this report the overall design, stratification and allocation plan, selection procedures, and estimation methods. Our budget consist of a total of \$500,000; each school C_n cost \$3,000 and each student C_m cost \$50 to sample.

The MDE is specifically interested in three outcome variables; every smoked one cigarette, every smoked marijuana, and age when first approach to smoke cigarettes or marijuana. Moreover, they provided us with expected levels of precision for the survey in terms of the expected means and coefficient of variation (CV=0.05) which can be seen in Table 1.

Table 1: Key Variables and Desired Levels of Precision

Outcome	Type	Desired CV	Expected Mean
smoked_cig	prop	0.05	0.25
smoked_mj	prop	0.05	0.15
age_approached	mean	0.05	12.00

Given the desired levels of precision, we can compute the desired simple random sample (SRS) sample sizes when CV =.05 as $n = \frac{s^2}{se^2}$. We first must calculate the element variance for each key variable. For proportions we use $\hat{p}(1-\hat{p})$, while for age we simply just square the estimated standard deviation of 1 that was given to us, $v(\bar{y}) = \sigma^2$. We then calculate the standard error as $se(\hat{p}) = CV \times \hat{p}$. Finally, we estimate the desired sampling variance as $var(\hat{p}) = se(\hat{p})^2$, where $se(\hat{p}) = \sqrt{var(\hat{p})}$. We show these results in Table 2, and note that these desired levels of precision would lead to large differences in sample sizes for each target variable. Therefore, we may wish to consider a more complex survey design.

Table 2: Estimating SRS Desired Sample Size

Outcome	Element Variance	SD	SE	Sampling Variables	SRS N
smoked_cig	0.1875	0.4330127	0.0125	0.0001563	1200
smoked_mj	0.1275	0.3570714	0.0075	0.0000562	2267
age_approached	1.0000	1.0000000	0.6000	0.3600000	3

2 Sampling Design

We propose a two-stage stratified cluster sampling as a complex survey sampling method for this project combining stratification with multi-stage cluster sampling. First, the entire population is divided into mutually exclusive and collectively exhausted subgroups called strata, based on shared characteristics relevant to the study (e.g., geographic region). Then, within each of these strata, the first stage of sampling occurs: a random sample of clusters (natural groupings like schools) is selected, and are called primary sampling unts (PSUs). In the second stage, a random sample of individual elements known as secondary sampling units (SSUs), such as students, are drawn from within each of the clusters selected in the

first stage. This approach aims to improve sample efficiency and representation by ensuring subgroups are included (i.e., stratification) while reducing costs and logistical challenges associated with sampling individuals across large geographic areas.

The MDE has given us the 2024 7th through 12th grade student headcount for each public and private school within each of the nine regions resulting in a target frame of $830,138$ across $2,443$ schools (78% Public). Schools in the first stage will be selected with probability proportional to student body size (PPeS). The proportionate allocation $M_h / \sum M_h$ where M_h is the total number of students in stratum h, will be used to determine school number selection in the first stage. Table 3 presents for each of the 9 strata total students, number of schools, and proportionate allocation.

Table 3: Proportionate Allocation Across Strata

Region	Total Student	Total Schools	Proportionate Allocation
1	3561	20	0.0042896
2	5474	30	0.0065941
3	8631	33	0.0103971
4	4855	31	0.0058484
5	18907	80	0.0227757
6	33133	133	0.0399126
7	191992	644	0.2312772
8	188830	549	0.2274682
9	374755	923	0.4514370

We obtained design effects (DEFF) estimates (DEFF_cig=2.5, DEFF_mj=2.0,

DEFF_age=1.7) from a similar pilot study of 7,500 students based on on 150 schools with 50 students each, and based on these we estimate the rate of homogeneity roh for each target variable. We use the provided DEFFs to estimate roh as $\hat{roh} = \frac{DEFF-1}{m-1}$, where m is the total sampled students in the pilot study, to consider alternative cluster sample designs along with cost-considerations. Table 4 provides the roh estimate for each target variable.

Table 4: We Use Pilot Study DEFFs to Estimate roh

Outcome	DEFF	roh
smoked_cig	2.5	0.0306122
smoked_mj	2.0	0.0204082
age_approached	1.7	0.0142857

2.1 *Sampling Within Budget*

Recall that our budget cost constraints as the cost per cluster as $c_n = \$3,000$ and cost per student as $c_m = \$50$, with a total budget constraint of $C = \$500,000$. We can use the rohestimates along with these costs to estimate the optimum subsample size m_{opt} needed to achieve the desired precision as $m_{opt} = \sqrt{\frac{c_n}{c_m}\frac{1-roh}{roh}}$. Note that since we have three target variables we also have three separate roh estimates and thus three m_{opt} estimates. Similarly, we can use m_{opt} to estimate the number of schools n_{opt} to sample, $n_{opt} = \frac{C}{c_n + m_{opt} \times c_m}$. Finally, we compute new DEFF for each variable as roh is portable and since we already computed m_{opt} as $DEFF_{new} = 1 + (m_{opt} - 1) \times roh$. By multiplying $m_{opt} \times n_{opt}$ for each variable we also get the total subsample size, as well as compute the total cost using $n_{opt} \times c_n + n_{opt} \times m_{opt} \times c_m$. Table 5 shows for each target variable m_{opt}, n_{opt}, $DEFF_{new}$,

total subsample size denoted by $total_{nm}$, and the total cost. Notice that our $DEFF_{new}$ estimates are close to those from the pilot study since again we used these to compute roh which is portable for estimating new design effects.

Table 5: Estimating New DEFF and Total Cost

Outcome	m_opt	n_opt	deff_new	total_nm	cost	Option
smoked_cig	43.58899	96.53536	2.303745	4207.879	$500,000	1
smoked_mj	53.66563	87.97734	2.074809	4721.360	$500,000	2
age_approached	64.34283	80.42281	1.904898	5174.631	$500,000	3

2.2 *Evaluating Alternative Clustering Designs*

We are left with a dilemma since we now have three subsample sizes that we can use, thus giving us three clustering design options to choose from. Using the m_{opt} values from table 5 as our three options, we iterate over each set of three target variables using these values to recompute for each target variable a new design effect and evaluate the estimated sampling variance for each design. We estimate the SRS sampling variance as $var_{srs} = \frac{var}{total_{nm}-1}$ where var is the sampling variance calculated from the pilot study and the denominator is the degrees of freedom. Additionally, we estimate the sampling variance for the clustering design as $var_{crs} = var_{srs} \times deff_{new}$. After estimating var_{crs}, we can square it to estimate a standard error, $se = \sqrt{var_{crs}}$ to use to estimate 95% confidence intervals for the estimated means. Additionally, in our evaluation of m_{opt} we can determine if the estimated sampling variance from the complex design is smaller than what is desired from MDE, therefore Table 6 shows sampling variances, standard errors, confidence intervals, and a variance check (e.g., "Y" = yes if <= to desired sampling variance) for each target variable using

the m_{opt} options in Table 5.

Table 6: Evaluating Alternative Clustering Designs

Outcome	Option	deff_new	var_srs	var_crs	se	lower	upper	var_ck
smoked_cig	1	2.303745	0.000057	0.000131	0.011464	0.227530	0.272470	Y
smoked_mj	1	1.869163	0.000030	0.000057	0.007527	0.135248	0.164752	N
age_approached	1	1.608414	0.002139	0.003441	0.058660	11.885027	12.114973	Y
smoked_cig	2	2.612213	0.000051	0.000133	0.011525	0.227412	0.272588	Y
smoked_mj	2	2.074809	0.000027	0.000056	0.007486	0.135327	0.164673	Y
age_approached	2	1.752366	0.001907	0.003341	0.057802	11.886707	12.113293	Y
smoked_cig	3	2.939066	0.000046	0.000136	0.011676	0.227114	0.272886	Y
smoked_mj	3	2.292711	0.000025	0.000057	0.007517	0.135267	0.164733	N
age_approached	3	1.904898	0.001740	0.003314	0.057565	11.887172	12.112828	Y

Upon evaluation of options from Table 6, we determine that option 2 has reasonable estimated design effects comparable to those from the pilot study as well as it is the only option to pass the estimated sampling variance check we designed. Option 2 with $m_{opt} = 53, n_{opt} = 88$ would cost a total of \$497,200. Since the total student population is 830138 and our now target sample is 4721 we can estimate the sampling fraction to be $f = n/N = 0.0056874$. The sampling fraction is the ratio of the sample size to the population size, and this estimate translates to the sample comprises approximately 0.57% of the total student population. In this case, the sampling fraction is low and the finite population correction factor is not needed for calculating variances to adjust for the fact that sampling without replacement from a finite population reduces variability compared to sampling from an infinite population.

2.3 *Non-response Adjustments*

The MDE anticipates 30% school response rates and 70% among students; therefore, we adjust the number of schools and within-school target by multiplying $m_{opt} \times RR_{student} = 53 \times .70 = 76.6651878$ students and for schools $n_{opt} \times RR_{schools} = \times .30 = 293.2578028$. We use these values to allocate the number of clusters for each strata based on the proportion allocated we calculated.

3 Stage 1 Selection

We consider stratified PPeS selection of schools from each strata by first sorting the list of schools to achieve implicit stratification. How we sort is by taking the number of 9th through 12th grade for each school divided by the total student body and descend order, and in this way we hypothesize that schools with older students are more likely to be positively associated with the target variables. For each strata h we assign our adjusted $n_{opt} \times proportionate\_allocation$ estimated earlier to calculate the number of schools to sample, denoted as n_h in Table 7. We use n_h to calculated in this Table the sampling interval $k_h = \frac{\sum_{i \ in h} MOS_{hi}}{n_h}$ where MOS_{hi} is the measure of size (MOS), total student head-count, for each school i in strata h. The k_h parameter is an important component of systematic sampling to determine how frequently units are selected from an ordered list. To conduct this selection, we randomly select a number between 1 and k_h for selecting schools from the list, which is captured in Table 7 as RN.

Table 7: Evaluating Alternative Clustering Designs

Region	prop_allocation	n_h	k_h	RN
1	0.0042896	1.257973	3561.000	3168
2	0.0065941	1.933767	2737.000	2310
3	0.0103971	3.049021	2877.000	1321
4	0.0058484	1.715096	2427.500	131
5	0.0227757	6.679161	2701.000	2122
6	0.0399126	11.704693	2761.083	2114
7	0.2312772	67.823846	2823.412	374
8	0.2274682	66.706826	2818.358	380
9	0.4514370	132.387420	2839.053	1673

For each strata we use the random start to select the first school, and for stratum with more than one selection we use $RN, RN+k_h, RN+2k_h, ..., RN+(n_h-1)$ until we satisfy n_h selection. Our minimum MOS 76.67 is also our m_{opt} and we use it here to determine the minimum number of students in each selected school required and if this is not satisfy we perform post-selection linkage. The linking is done by first selecting the number of schools in each strata. When the next units on the list do not meet the sufficient MOS size required we move forward in the list until the first unit that meets the minimum requirement is achieved. For all the units that did not meet the requirement they are cumulated backwards until a linked unit of minimum sufficient size is created. We do this process for all strata. Table 8 shows for each region the total number of clusters, total schools linked, and total student count.

Region	k_h	RN	m_h_star
1	3561.000	3168	20.25297
2	2737.000	2310	15.56652
3	2877.000	1321	16.36276
4	2427.500	131	13.80626
5	2701.000	2122	15.36177
6	2761.083	2114	15.70349
7	2823.412	374	16.05798
8	2818.358	380	16.02924
9	2839.053	1673	16.14694

3.1 *Pseudo Strata Method*

The PPeS systematic sampling makes for direct variance estimation difficult; therefore, we use the paired selection model to help us estimate variance. However, we cannot form a pseudo-stratum from just one cluster nor have odd number of clusters, since paired methods require at least tow units within a stratum. A key constraint for this approach is that each stratum must contain at least two independent variance units. We collapse stratum that have odd number clusters or just one cluster with the adjacent stratum. In Table 9 we show collapsed strata and corresponding number of sampling error computation units (SECU); we collapsed regions 1 to 3 into stratum 1, regions 4 to 7 into stratum 2, and left regions 8 and 9 the same. We take all of the linked clusters and divide them by two to get the paired

SECU count. We now have a total of 4 strata in this model, 3 SECUs for stratum 1, 103 SECUS for stratum 2, 77 SECUS for stratum 3, and 135 SECUs for stratum 4 resulting in a total of 318 SECUs across strata.

Table 9: Evaluating Alternative Clustering Designs

Psuedo	SECU
1	3
2	103
3	77
4	135

Students selection sample

We implement the same systematic random sampling for the roster example of schools from Region 7. The randomly sampled middle school was from Region 7, the MOS for this school was 242, but the actual size is 219. This is the formula for calculating overall sampling fraction:

$$f_h = \frac{n_h MOS_{hi}}{\sum_{,i\in h} MOShi} \frac{m_h^*}{MOS_{hi}}$$

We got the expected sample size of 14.53126, and we rounded it to 15. To obtain the expected sample size, we first got the second-stage sampling rate is Sampling Rate $= \frac{m_7^*}{MOS_7} = \frac{16.05737}{242} = 0.06635277$, and then multiplied the rate by the actual sample size.

To get the sampling interval $k_h i = Mos_{hi}/n_h = \frac{16}{242}$, we choose a random starting number between 1 and k_{hi}, which is 14.6, then we use the k-interval 146 to conduct the systematic sampling. Then we selected the student at the random start position(14) and every k_{hi}-th student thereafter from the ordered roster. We print the roaster of names in the Appendix

to save space here.

3.2 *Sampling Error Calculation Model*

Our design resulted in 318 paired units of PSUs, used for variance estimation. These SECU pairs were formed by grouping sampled schools within collapsed regions to meet minimum size thresholds. The paired difference method requires at least two SECUs per variance stratum, and collapsing ensured that this condition was met in all cases. The degrees of freedom for variance estimation are calculated as the number of SECUs minus the number of collapsed strata, $df = a - H = 318 - 4 = 314$.

To estimate the variance of a ratio estimator (e.g., the proportion of students who smoked), we applied Taylor Series Linearization with paired SECUs. For n_h = 2 in each stratum, the variance of the ratio $= \frac{t_y}{/} t_x$ was approximated as:

$$\begin{aligned}\text{var}(r) \approx \frac{1}{\hat{t}_x^2}\left[\sum_h \text{var}(\hat{t}_{h,y}) + r^2 \sum_h \text{var}(\hat{t}_{h,x}) - 2r \sum_h \text{cov}(\hat{t}_{h,y}, \hat{t}_{h,x})\right] \\ = \frac{1}{\hat{t}_x^2}\Bigg[\sum_h (1-f_h)(\hat{t}_{h,1,y} - \hat{t}_{h,2,y})^2 + \\ r^2 \sum_h (1-f_h)(\hat{t}_{h,1,x} - \hat{t}_{h,2,x})^2 - \\ 2r \sum_h (1-f_h)(\hat{t}_{h,1,y} - \hat{t}_{h,2,y})(\hat{t}_{h,1,x} - \hat{t}_{h,2,x})\Bigg]\end{aligned}$$

In our case, the estimated variance across SECU pairs was var_estimate = 0.00514. From this estimate we estimate the standard error (SE) was as $se = \sqrt{0.00514} = 0.07179$. With degrees of freedom $df = 314$, the 95% confidence interval was calculated using the t-distribution: $(0.0588, 0.3403)$.

3.2.1 *Estimator Formula and Weighting*

The estimator for the proportion of students with a given trait (e.g., smoking behavior) is computed as a weighted average. The formula used is $\hat{p} = \frac{\sum w_{hij} y_{hij}}{w_{hij}}$. In this equation,$ y_{hij} = 1$ if student j in school i within stratum h has the characteristic of interest, and $y_{hij} = 0$ otherwise. The term w_{hij} represents the final weight assigned to each student and reflects the product of several adjustments. These include the inverse of the probability of selection, school-level and student-level nonresponse adjustments, and any corrections for second-stage linking or operational deviations during data collection.

Although the original design was structured to be equal probability of selection (epsem), in practice, weights are required. This is due to observed differences in participation and inclusion at both the school and student levels. Specifically, weights correct for approximately 30% nonresponse at the school level and 70% nonresponse at the student level. Thus, while our estimator maintains design-based integrity, proper weighting is essential for unbiased inference.

3.2.2 *Subclass Estimation*

The client requested reliable estimation for a 20% population subgroup. To evaluate the adequacy of our design for supporting such subclass estimates, we calculated the expected number of completes per SECU, denoted b*, which was approximately 76.67. Assuming a 20% subclass size, the expected number of observations per SECU for the subgroup is the expected subclass per SECU = 0.20 $times$76.68 = 15.3. This value exceeds the minimum recommended threshold of 10, which is generally considered sufficient for reliable variance estimation at the subgroup level. Therefore, we conclude that confidence intervals for subclass means or proportions can be computed using the same paired SECU difference

method, and the degrees of freedom remain the same.

4 Appendix 1: Students Selected From One School.

Table 10: Evaluating Alternative Clustering Designs

ID	Grade	class	Firstname	Lastname
1	7	Grady Vest	Magee	Monica L
16	7	Grady Vest	Schwartz	David Scott
30	7	Qixuan Li	Raynor	Gregory K
45	7	Qixuan Li	Franke	Mira A
59	7	Bill Pesau	Black	Stephen P
74	7	Andrew Bellman	Marbut	Joanne Renee
89	7	Andrew Bellman	Trecker	Molly A
103	8	Joe Williams	Dawson	Rebecca S
118	8	Joe Williams	Krosky	Paula Michele
132	8	Robert Mcfay	O Brien	Erin Terese
147	8	Joseph Miden	Kosove	Daniel Brian
162	8	Joseph Miden	Loevy	Debra L
176	8	James Vogner	Dworzanowski	Gregory William
191	8	James Vogner	Gatto	Julia Lynn
205	8	Jane Doe	Kaye	Peter Mitchell