Michigan Teen Smoking and Drug Use Survey Sample Design

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

April 25, 2025

Abstract

The State of Michigan Department of Education (MDE) requires data to monitor teenage smoking and drug use to assess compliance with tobacco industry settlements. This report outlines a two-stage stratified cluster sampling design developed for the MDE to monitor smoking and drug use among students in grades 7-12 statewide. The design addresses the need for cost-effective, statistically sound estimates at both the state and regional levels, meeting specified precision targets (CV=0.05) within a $500,000 budget. Schools were stratified into nine regions and selected with probability proportional to size (PPeS), followed by systematic selection of students within schools. The final design targets approximately 88 schools and 4,721 students, incorporating adjustments for non-response and procedures for handling undersized schools to maintain equal selection probabilities. The report details allocation, selection methods, and an estimation plan using a paired selection model for variance calculation.

1 Introduction

The State of Michigan Department of Education (MDE) is specifically interested in three outcome variables; ever smoked one cigarette, every smoked marijuana, and age when first approach to smoke cigarettes or marijuana. Moreover, MDE provided us with expected levels of precision means and coefficient of variation (CV=0.05) shown 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

From 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 square the estimated standard deviation of 1, $v(\bar{y}) = \sigma^2$. 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 Appendix 1, 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.

2 Sampling Design

We employ a two-stage stratified cluster sampling design as it 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. In the first stage, random selection of schools (e.g., Primary Sampling Units [PSU]) are determined by proportionate allocation for each strata. The second stage randomly selects students (Secondary Selection Units [SSU]) from the selected school clusters within each region.

The MDE provided 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. Appendix 2 presents for each of the 9 strata total students, number of schools, and proportionate allocation.

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 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. Appendix 3 provides the roh estimate for each target variable.

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 2 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 2: 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 have the possability of three clustering design options to choose from based on the target variables. Using the m_{opt} values from Table 2 as our three options, we iterate over each set of 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, Appendix 4 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.

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 dividing$m_{opt} \times RR_{student} = 53.6766 \div .70 = 76.6651878$ students and for schools $n_{opt} \times RR_{schools} = 87.9689 \div .30 = 293.2578028$. 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. We sort by taking the number of 9th through 12th grade for each school divided by the total student body and descending 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 Appendix 5. We use n_h to calculated in this Table the sampling interval $k_h = \frac{\sum_{i\ inh} MOS_{hi}}{n_h}$ 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. We randomly select a number between 1 and k_h for selecting schools from the list, denoted as RN.

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 satisfied 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 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. Appendix 6 shows for each strata the total number of clusters, how many totaled schools linked and the total number of students.

4 Stage 2 Selection

We assume rosters are made available by the school administration at the time of data collection. These rosters are ordered and formatted uniformly to facilitate systematic sampling. To maintain equal probability of selection (epsem) across all strata, we computed the required number of students to be sampled per selected school, denoted as m_h^* , based on the within-strata sampling fraction f_h, which is the same as the overall sampling fraction f for all h, and the stratum-specific PPS sampling interval k_h as follows, $f_h = \frac{n_h MOS_{hi}}{\sum_{i \in h} MOShi} \frac{m_h^*}{MOS_{hi}} = \frac{n_h m_h^*}{\sum_{e \in h} MOS_{hi}} \Rightarrow m_h^* = f \times k_h$. This ensures that when each school is selected with probability proportional to its MOS and then students are sampled within school at a fixed rate $\frac{m_h^*}{MOS_{hi}}$, the overall inclusion probability for any student is, $\pi_i = \frac{n_h \times MOS_{hi}}{\sum MOS_{hi}} \times \frac{m_h^*}{MOS_{hi}} = \frac{n_h m_h^*}{\sum MOS_h}$. Each student has the same selection probability within the state and within region h, satisfying the epsem condition. Table 3 summarizes the first stage stratification and selection for two strata. The tolerated minimum number of students per school is estimated as m_h^* divided by the expected student response rate of 0.70 and is denoted as b_h in the table. Additionally, we compute the sampling interval for each stratum to achieve epsem, as well as show the random start used to select schools.

Table 3: Sampling Interval for Two Strata,

Region	f_h	MOS_h	n_h	b_h	k_h	RN
3	0.0321459	1552	3.049021	23.37538	2877.0	1321
4	0.0306724	772	1.715096	19.72323	2427.5	131

4.1 *Undersized Schools Linking*

In cases where a selected school had fewer students than the desired cluster size m_h^*, the within-school sampling rate would exceed 1.0, making the design unfeasible. To address this, and to ensured that the effective number of completed questionnaires per school meets the targets m_h^*, we linked undersized schools with nearby schools when their MOS was less than $\frac{m_h^*}{r}$, where r is the expected student response rate (70%). This operational rule preserves feasibility without altering theoretical inclusion probabilities. The within-school sampling rate remains $\frac{m_h^*}{MOS_{hi}}$, maintaining epsem across students. For example, 5 small schools in Region 4 were linked to form a cluster of 10 meeting the required sample size of 19.729. Note that no oversize schools were identified that required splitting, and all selected schools acceptable size or linked as needed. Appendix 7 presents the within school sampling interval for the two selected strata in Table 3.

4.2 Student 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 14.5313, 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 $k_{hi} = MOS_{hi}/m_h^* = \frac{219}{15}$, 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. The roaster of names is in Appendix 8.

5 Estimation Plan

5.1 *Pseudo Strata*

We propose using the paired difference method and create pseudo-strata to estimate variance by using Taylor Series Linearization. However, since we cannot form a pseudo-stratum from one cluster nor have an odd number of clusters, each pseudo-stratum requires at least two units. We combine strata that have odd-number clusters or just one cluster with the adjacent stratum. We collapse region 1 to 3, then we randomly group odd and even selections of units in the stratum into two sampling error computation units (SECUs). For regions 4 to 7 we also collapse them, then group them into two SECUs. For region 8 and region 9, we keep them the same, each with 2 SECUs. Thus, we get 8 pseudo-strata in total.

5.1.1 *Variance Estimation*

In this study, we need to estimate the proportion who have ever smoked, the proportion who have ever used marijuana, and the mean age at first use of cigarettes or marijuana. We did not consider using weight here since we maintain epsem design through the whole sampling process, which means every student has the same probability of being selected.However, in practice, we need to consider the response rate (e.g., the response rate among schools will be 30 percent, and the response rate among teenagers within schools will be 70 percent), which we should adjust our weight in design based on the response rate when conducting variance estimation. We decided to use Taylor Series Linearization to estimate the ratio estimator, which is approximated as For all strata, $n_h = 2$.

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

The general estimator used is the ratio estimator: $\hat{r} = \frac{\hat{t}_y}{\hat{t}_x}$

- $\hat{t}_y$: estimated total for the numerator variable(e.g., number of students who smoked)
- $\hat{t}_x$: estimated total for the denominator(e.g., total eligible students)
- $\text{var}(\hat{t}_{h,y})$: variance of the numerator total within stratum h
- $\text{var}(\hat{t}_{h,x})$:variance of the denominator total within stratum h
- $\text{cov}(\hat{t}h, y, \hat{t}h, x)$: covariance between numerator and denominator totals within stratum h

5.1.2 *Confidence Interval*

A 95% confidence interval for the estimated proportion and mean $\hat{r}$ is given by: $\hat{r} \pm t_{df,0.975} \cdot \text{SE}(\hat{r})$, where, $SE(\hat{r}) = \sqrt{\text{Var}(\hat{r})}$, and the df is equal to the number of pseudo strata which is 8. We conducted a simulation of the variance estimation of the proportion of students who ever smoked by applying the variance estimation formula, and we estimate the ratio to be 0.249946, SE = 0.000055, 95% confidence interval [0.249815, 0.250069].

5.1.3 *Subclass Estimation*

For the 20% subgroup, we used the same estimation and variance formulas, but applied them to a subset of the data that reflects 20% of the full population(lower-income households). However, some SECUs may contain a very small number of students. The expected sample size for this subgroup is 944.2 (total sample size $\times prop = 4721 \times 0.2 = 944.2$), and the expected subclass size per cluster is 15.33304 ($E[b^{*}] \times subclass_{pct} = 76.66519 \times 0.20 = 15.33304$). Besides, some strata contain only one SECU, and which cannot allow for variance estimation. If too few SECUs contribute, degrees of freedom could be too low, affecting reliability. Low-income students might not be spread evenly across all PSUs, and if only a small number of students in the subclass are sampled per stratum, variance estimates will be unstable or undefined. Therefore, the design might not accommodate accurate inference for the 20% subclass group.

In conclusion, this report details a robust and statistically efficient two-stage stratified cluster sampling design tailored to the MDE's need for monitoring teenage smoking and drug use. By employing stratification across educational regions, probability proportional to size selection for schools, and systematic sampling of students within schools, the design ensures representative statewide and regional estimates while adhering to budget constraints. The chosen parameters, including an anticipated sample of approximately 88 schools and 4,721 students, are optimized to achieve the required precision for key variables after accounting for anticipated non-response. The outlined procedures for selection, including linkage for smaller schools, and the comprehensive estimation plan provide a clear roadmap for survey implementation and analysis, ultimately delivering a cost-effective solution capable of generating the critical data required by the MDE.

6 Appendix 1: Estimating SRS Desired Sample Size

For each target variable alongside desired levels of means and sampling variance we display below the element variance, standard deviation and standard error, and sampling variance. We use the sampling variance as the denominator to determine SRS sampling size for each target variable.

Outcome	Element Variance	SD	SE	Variance	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

7 Appendix 2: Proportionate Allocation Across Strata

We compute a proportionate allocation of students across all nine strata. For instance, 45% of students across 923 schools in this population come from stratum 9; therefore, the proportionate allocation for this stratum is .4514 in the below table.

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

8 Appendix 3: We Use Pilot Study DEFFs to Estimate roh

We estimate roh form the design effects provided from a similar pilot study. Below we show roh estimates for each target variable.

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

9 Appendix 4: Evaluating Alternative Clustering Designs

The table below presents the three optimum subsample sizes for each target variable alongside estimates of sampling variance, standard errors, and a variance check to determine which option is optimal for this design.

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

10 Appendix 5: Strata Cluster Selection, Sampling Interval

We calculate using the proportionate allocation and n_opt the number of clusters to select from each strata. Additionally, we compute sampling intervals and select random starts from 1 to k.

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

11 Appendix 6: Linkage of Schools for Stage 1 Selection

We show the number of clusters to be sampled for each strata along with the number of total linked schools and total number of students in each strata to use for random selection in the second stage.

Region	Clusters	Schools	MOS
1	1	1	430
2	2	2	933
3	3	3	1552
4	10	14	772
5	7	7	4400
6	12	12	8202
7	177	281	52360
8	154	224	59827
9	270	379	141443

12 Appendix 7: Within-School Sampling Interval for 2 Strata

We show the selection of schools below after selection for two strata and within-school sampling rate. The within-school sampling interval was calculated as the total student count divided by the target sample size (m_h_star). Students were then selected systematically using a random start between 1 and the interval.

Selection Num.	Stratum	School	Cumul MOS	Within School Interval
4	3	01155	928	56.714138
5	3	01527	1355	26.095837
6	3	04860	1552	12.039531
7	4	02692	323	23.395186
8	4	06812	387	4.635579
9	4	08446	444	4.128562
10	4	08063	493	3.549115
11	4	03998	540	3.404253
12	4	04034	586	3.331822
13	4	09308	629	3.114529
14	4	02305	670	2.969668
15	4	08521	703	2.390220
16	4	07124	723	4.997733

13 Appendix 8: Students Selected From One School.

Table 11: 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