Manual of the China Gen-Z Health Behavior Survey 2022-23

Grant source: CMB21-436

Principal Investigator: Xiaozhao Yousef Yang

Sample Design and Sampling

Selection of ideal typical regions

The Chinese Gen-Z Health Behavior Survey 2022-23 (hereafter referred to as the survey) adopted a multiple stage complex sampling combining quota sampling and cluster sampling at the SSU-level. The largest sampling frame of China was divided into four compartments with an intent to qualitatively saturate the ideal types of Chinese regions. This consideration was motivated the insights into the cultural diversity and homogeneity of the country and the epidemic status of electronic cigarette use. The survey intentionally left aside the regions that are culturally heterogenous with a non-Sinitic heritage, including Tibet, Xinjiang, Manchuria, Inner Mongolia, and the Muslim belt of northwestern China.

As a result, the survey selected four ideal typical regions: the Yangtze Delta, the Pearl Delta, northern inland, and southern inland. These four regions constitute the contour of China proper east to the Hu Line, where over 90% of the population reside. The two Deltas were chosen and over-sampled because e-cigarette use and its market share in China first penetrate into major metropolis and the more developed coastal urban areas. The two Deltas are the most developed and modernized regions of China sitting on eastern and southern coasts. The indigenous population of Yangtze Delta speak Wu and follow the distinct Wu-yue culture. The

indigenous population of Pearl Delta are culturally Nam-yue residents who speak Cantonese. Meanwhile, inland China has a relatively lower level of socioeconomic development and more traditional organization of social structure. The sample from northern inland draws from the Mandarin-speaking population, whereas the sample from southern inland mainly targets the speakers of Southwestern Mandarin. One should note that a significant number of respondents were from migrant families or were out-of-state students, hence the survey used cultural ideal types to characterize the sampled regions, not the individual respondents.

Because each region is an ideal type of a distinct Sinitic culture, in order to minimize the internal variance within each region, the choice of municipals in each region was also motivated by their classical cultural ideal-typicality rather than a probability-based frame. However, the survey imposed a quota in each region that requires at least one municipality to sample rural districts exclusively. In Yangtze Delta, four municipalities within the koine Taihu Wu cultural sphere were selected: Shanghai (urban), Suzhou (urban), Hangzhou (urban and rural), Yuyao (rural). In Pearl Delta, three municipalities that traditionally speak the standard Gwongfu Cantonese were selected: Guangzhou (urban), Shenzhen (urban), Zhongshan (rural). In southern inland, three municipalities were selected: Chongqing (urban), Chengdu (urban and rural), Kunming (rural). In northern inland, three municipalities were selected: Beijing (urban), Taiyuan (urban), Longkow (rural).

Selection of schools and classes

Because the four regions were selected based on cultural ideal types instead of a probability-based frame, so we may not consider the four regions or the chosen municipalities therein as

PSUs. Instead, the sub-municipal districts were treated as the PSUs for probability-based cluster sampling. Within each PSU, the SSUs were schools. The survey imposed a minimum quota so that each region will sample each of the four types of SSU: high school, vocational high school, college, vocational college. Next, within each SSU, about two classes of similar sizes were sampled as tertiary sampling units. This is determined by $n = z_{a/2}^2 v/e^2$ to achieve a desired CI of 95%, with variance=0.2, e^2=0.2, n≈2. Questionnaires were administered to all consenting students in the tertiary sampling units.

The survey invited identified school coordinators from all four types of schools in each region. Once agreed to participate, the survey requested two to three classes to participate. Invitation was sent through email, WeChat, phone, and personal communication. A second follow-up was forwarded one week after the first invitation. Advance letters in Chinese were sent to eligible school coordinates and teachers explaining the purpose, ethics approval, fund source, and compensation of the survey. Non-response rate originated mainly from two stages: PSU not participating, insufficient SSUs participating. Non-response rate from PSUs and SSUs were estimated in Table 2.

At the level of individual respondent, non-response rate was very low. The students were informed with consent statement and were free to withdraw at any stage during the self-administered survey. The survey deployed several strategies to minimize non-response rate at the individual-level. First, the survey was distributed by school coordinators and teachers in class. Students were also able to complete the survey at another time at their convenience. Second, most questionnaires were distributed electronically via a secured link. Besides the informed consent statement, the heading of the questionnaire informed the students that their

teachers cannot access to any information collected in the questionnaire and of how the data will be stored securely. Third, hand-writing was minimized in paper questionnaire with only one item that required hand-writing fill-in.

After the survey, due to the lower participation from Inland South, the survey further invited one school from urban Chengdu to participate. The survey also reinforced the Inland South sample with 100 respondents from Chongqing who were sampled by an outsource survey company.

Table 1. Characteristics of the four regions

Ideal-typical regions	Sampled municipalities	Culture	Main language	GDP per capita in 2021	Population 2020
Yangtze Delta	Shanghai, Suzhou, Hangzhou, Yuyao	Wu	Wu-Taihu dialect	$18k	39.4Mil
Pearl Delta	Guangzhou, Shenzhen, Zhongshan	Cantonese	Cantonese-Gwongfu dialect	$19.7K	28.3Mil
Southern Inland	Chongqing, Chengdu, Kunming	Basu, Yunnan	Southwestern Mandarin	$9.8k	36.8Mil
Northern Inland	Beijing, Taiyuan, Longkou	Northern Chinese	Northern Mandarin	$12.6k	23.1Mil

Figure 1. Geographical distribution of the sampled municipalities

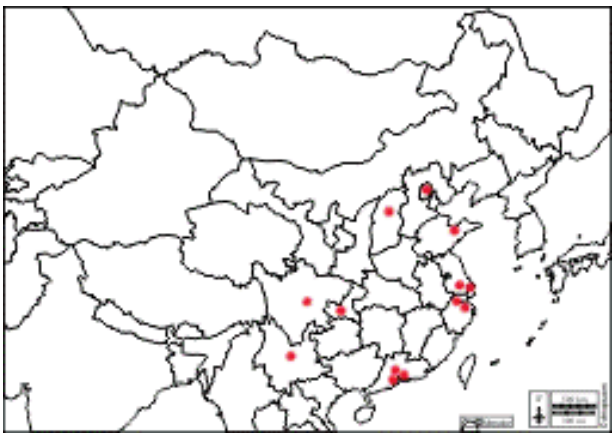

Table 2. Response rates

Ideal-typical regions	Municipalities	School (attempted vs. obtained)	Response rate	Class (needed vs. obtained)	Response rate
Yangtze Delta	Shanghai	1,1	100%	2,2	100%
	Suzhou	2,2	100%	4,5	100%
	Hangzhou	7,5	72%	10,12	100%
	Yuyao	2,2	100%	4,6	100%
Pearl Delta	Guangzhou	9,7	77.8%	14,14	100%
	Shenzhen	2,2	100%	4, 5	100%
	Zhongshan	2,1	50%	2, 2	100%
Southern Inland	Chongqing,	3,2	66.7%	4, 3	75%
	Chengdu,	6,4	66.7%	10, 7	70%
	Kunming	3,2	66.7%	2, 2	100%
Northern Inland	Beijing	3,2	66.7%	4, 3	75%
	Taiyuan	6,4	66.7%	6, 6	100%
	Longkou	2,2	100%	4, 2	50%

Multi-stage complex sampling design

Sample sizes of PSU and SSU

Because the variance of standard school size is relatively small, the number of tertiary sampling units (classes) were determined *a priori* to be equal. Therefore, the survey's design is reduced to

two-stage cluster sampling with a quota on school type. In two-stage cluster sampling, the goal of the survey is to cover as much variability with as least cost and inconvenience. At the first stage of choosing PSU, treating all PSU as equal without replacement, simply use formula (Cochran 1977):

$$m = \frac{z_{\frac{a}{2}}^2 v}{e^2} \times \frac{M}{M-1} + z_{a/2}^2 v/e^2$$

Where $z_{\frac{a}{2}}^2$ is 1.96 for 5% of significance level, v/e^2 is .10*.90 for the default level of 10% smoking prevalence from our pilot study in Guangdong province, m is estimated in column 4 of Table 3.

For the size of SSU in each PSU, we consider the internal and inter-PSU variability of SSU to determine the desired sample size. One approach (Lohr, 2010) is to minimize $V(\hat{\bar{y}}) = \frac{\left(1-\frac{n}{N}\right)MSB}{nM} + \frac{\left(1-\frac{m}{M}\right)MSW}{nm}$, where MSB and MSW are the mean squares between and within PSU. Consider the simple cost function as $C = c_1 n + c_2 nm$, where n is the sample size of PSU and m is that of SSU, optimal value:

$$n_{opt} = \frac{C}{c_1 + c_2 m_{opt}}$$

$$m_{opt} = \frac{\sqrt{c_1 M(N-1)(1-R_a^2)}}{\sqrt{c_2}(NM-1)R_a^2}$$

$$R_a^2 = 1 - MSW/S^2$$

Once a school participates, classes follow suit without much difficulty. Thus, the cost of sampling PSU is higher than that of sampling SSU, set c_1=$4 * c_2$. The ratio of 4 is also the amount the survey

compensates school coordinators versus teachers. Assuming MSW is 70% of the total variance, the homogeneity measure R_a^2 is 0.3, m_{opt} is .88 for the first three regions and .90 for Inland North. The number of SSU in PSU_i is N_i, the average N in each regional sample frame is $\sum_{i=1}^{M} N_i \,/M$. Optimal PSU and SSU sizes can be determined accordingly and are listed in Table 3. In each region, optimal PSU indicates the number of SSU needed in each PSU that both maximizes variability and minimizes costs. The products of optimal PSU value and optimal SSU are shown in column 5.

Probability weighting and design effect

In general, each standard class has 40 seats, the required respondent numbers (R) are listed in column 6 of Table 3. Actual sampled numbers of PSU, SSU, and respondents, along with their weights (w) are listed in columns 1-3 of the right side panel of Table 3. These weights allow researchers to calculate the unequal weighting effect (UWE) and design effect (Deff):

$$UWE = 1 + s_w^2/\overline{w}$$

$$Deff = \sqrt{UWE + (M-1)R_a^2}$$

where $\overline{w}$ is the average of weights for all units and s_w^2 is their variance. UWE and Deff are displayed in column 4-5 of the rightside panel of Table 3.

Table 3. Multi-stage complex survey design effects and weights

Region	#PSU (M)	#SSU each PSU (approx.)	m	$n_{opt} * m_{opt}$	R_{opt}	m, w	n, w	n2, w	UWE	Deff
Yangtze Delta	34	14.0	17.4	17.4*.88=15	600	8, 2.22	10, 1.2	1156, .52	1.44	3.37

Pearl Delta	26	15.1	17.6	17.6*.88≈15	600	7, 2.93	10, 1.2	689, .87	1.45	3.01
Inland South	35	15.2	17.4	17.4*.88≈15	600	5, 3.48	6, 2.0	439, 1.36	1.20	3.37
Inland North	25	23.1	17.6	17.6*.90≈16	640	6, 2.51	10, 1.2	587, 1.02	1.41	2.93

Surveys with multistage complex design will sample units with unequal probability and thus require certain weighting strategies for sampling units at different levels. Ignoring weights will result in incorrect estimates and type I error in hypothesis testing. Besides adopting weights, researchers often adjust unequal probability and artificial variance resulted from clustering by using robust errors bounded by clusters or other design-based analysis. At times, but with great caution, researchers also use model-based adjustment that corrects the probability of inclusion by controlling certain background covariates that may be related to sampling units being selected, such as digital literacy, distance to school, etc.

Sampling weights that adjust for the unequal selection probability resulted from survey design features are listed in the right-side panel of Table 2. These are the inverse of the probability that a specific PSU, SSU, and respondent was included in for the survey. Users may employ the weights in the *survey, [pw], [aw]* functions of Stata depending on their need. If users wish to apply post-stratification weights based on the difference between the distribution of demographic characteristics in this sample and that in the population, they may consult particular statistical yearbook or official release for the population information from the specific municipalities that were chosen in this survey.

To estimate total, for example, first use unbiased estimator for individual in *i*th SSU: $\widehat{t_i} = \sum_{j \in S_i} \frac{y_{ij}}{\pi_{i|j}}$. Then use Horvitz-Thompson estimator to get population total by indicating whether SSUs are in the sample: $\hat{t}_{HT} = \sum_i^N Z_i \frac{\hat{t}_i}{\psi_i}$, where Zi=1 if psu i is in the sample, ψ_i is the probability that psu i was selected.

Once obtained the estimated total, the population mean is total divided by weight $\hat{\bar{y}}_{HT} = \frac{\widehat{t_i}}{\sum_{k \in S} \sum_{j \in S_k} \sum_{i \in S_j} w_{ijk}}$. The residuals from the estimated psu totals are $\hat{e}_i = \hat{t}_i - \hat{\bar{y}}_{HT} \widehat{M}_i$, we replace the total number of ssus in the population M0 for $\widehat{M}_i$ for convenience. The estimated variance of the mean is then:

$$\hat{V}_{WR}(\hat{\bar{y}}_{HT}) = \frac{n}{n-1} \sum_{i \in S} (\frac{\hat{e}_i}{M_0 \pi_i})^2 = \frac{n}{n-1} \sum_{i \in S} (\frac{\sum_{j \in S_k} \sum_{i \in S_j} w_{ijk} (y_{ijk} - \hat{\bar{y}}_{HT})}{\sum_{k \in S} \sum_{j \in S_k} \sum_{i \in S_j} w_{ijk}})^2$$

Standard error is the root squared variance, due to design effects, the 95% CI is $\hat{\bar{y}}_{HT} \pm 1.96 \sqrt{deff} \sqrt{\frac{\hat{V}_{WR}(\hat{\bar{y}}_{HT})}{n}}$.

Field Procedures

Pretest

A pretest study was initially conducted in two schools in Guangdong to test the completion time, questionnaire item orders, sensitivity, completion rate, semantic validity via electronic questionnaires. A separate earlier project (n=727) using stratified sampling across the 21 municipals of Guangdong province on the young adult population was conducted in early 2022

with a majority of similar questions that were later reused in the current survey. The pre-test survey was retrieved and revised per the feedbacks on semantic understanding, face validity, and cognitive difficulty of the responses. A panel consisted of three faculty members from three different institutions and six graduate students was invited and involved in the revision of the questionnaire based on the pre-test results.

Interviewer training

The survey was self-administered by students attending the class or allowing them to complete at any time at home. The local coordinators at each school informed the students of the nature and confidentiality of the survey, before distributing the paper-form questionnaires (in minority cases) or the link to the encrypted online survey platform (in majority cases). The coordinators were instructed to orally or in written form remind the respondents of the survey twice if completion rate was lower than 50% in each instance. A letter to respondents and letter to coordinators were both made available at each participating school, the content of which conveyed the nature, confidentiality, anonymity policy, source of funding, and purposes of the survey. Communication with coordinators and interviewers were made timely accessible via the survey teams' cell phones and emails, in order to offer ongoing support and guidance throughout the survey administration process.

Incentives

In the letter to coordinators, the compensation and incentive policy to participating coordinators and teachers detailed: for each participating class with more than 30 students and fewer than 60 students, 800 Yuan will be paid to the coordinator/teacher for each cohort of 25

returned questionnaires. Each additional cohort in the size of 25 will be compensated by 500 Yuan. The coordinators reserve the arbitration about how to compensate the students. In many cases the coordinators refused the monetary compensation citing good faith, friendship, devotion to research, so we have compensated them out of good faith by gifting token food such as several pounds of beef flank and beef meatballs.

Screening and consent

The survey design was approved by the ethics committee of the School of Sociology and Anthropology at Sun Yat-sen University. Funding comes from the US-China Medical Board (#21-436). The respondents were informed with consent to withdraw at any stage of the survey. The instruction at the beginning of the questionnaire, in both paper-form and electronic form, described the funding source and purpose of the study, as well as how long it will take (between 6 to 12 minutes based on the pre-test). The respondents were also ensured that their teacher coordinators will not access the data collected by the survey. Once the respondents "agreed" to participating, they will proceed to the next page of the questionnaire. In the electronic survey, five questions at the very beginning are required to be completed before proceeding to the next, they are the basic demographic questions on age, gender, ethnicity, estimated academic ranking, and relationship status. These required questions can be used in the imputation of missing data should some respondents leave a large chunk of the survey unfilled.