Abt's Proposed Approach to Sampling for the Survey of Respirator Use and Practices (SRUP)

August 23, 2023

Overall

Research Goal: determine respirator use by industry type and employment size category.

The target population consists of companies that have operations in the USA, regardless of whether they have their headquarters in the USA. Included in the sample frame are companies in all NAICS codes except for Public Administration (NAICS code 92). The study will include both companies that don't use respirators and those that do so that company behavior during the COVID Pandemic can be understood. Including both companies that do and do not use respirators in the sample will also allow Abt Associates to model the prevalence of respirator use across different industries.

The sample frame for this study would be a targeted listing of companies with a single location or the company headquarters. Companies with multiple locations will be represented by the information provided by the company headquarters staff. Subsidiaries will either represent themselves or will be covered by reporting from their holding company.

An unbiased sample frame includes a listing of all the entities of a target population. The most authoritative information about US companies is managed by the Bureau of Labor Statistics (BLS). This information is not accessible by Abt Associates or NIOSH. As a result, Abt does not have the ability to construct a comprehensive frame that contains contact information for all entities that are members of the target population. In addition, any other source of information about US companies will likely contain entities that do not belong in the frame. This creates a need to explore other options, such as adaptive sampling, for obtaining a representative sample of the target population.

Dun & Bradstreet (D&B) offers a comprehensive listing of companies in the USA, but Abt believes that there are limitations to the D&B database that make its coverage different than what BLS and the Census Bureau use. BLS has access to government resources such as tax records, regular industry and employment surveys, and information from other agencies such as the Energy Information Administration (EIA), the U.S. Department of Agriculture (USDA), and the Tennessee Valley Authority (TVA) that they have used to develop data about US companies. D&B uses company credit information that they have collected through their D&B Data Universal Numbering System (DUNS). D&B then further enhances this information with public data sources to develop and maintain their database. As the D&B data source is not directly aligned with the Census Bureau industry information, the possibility of under and over coverage exists. As a result, the possibility of frame errors and biases exists. Similarly, other frame-based sampling solutions Abt has explored would have similar issues and thus could also introduce frame error.

Abt proposes using an adaptive sampling scheme to identify members of the target population and, if they are valid members of the target population, to determine where respirator use takes place.

The recommended sampling strategy will be an iterative process that focuses on identifying and sampling firms that are part of the target population in three stages. Abt recommends a multi-stage adaptive sampling approach where information from each sampling wave informs the composition of

subsequent sampling waves. Specifically, Abt will use the first wave to help select likely entities for the next wave. In addition, this information will be used to adjust the sampling protocols to avoid surveying entities that are not part of our target population and obtain sufficient sample size across subgroups of the target population.

Sampling strategy

Terminology – Formula variables and definitions

North American Industry Classification System (NAICS) codes	
NAICS 2-Digit Code	*Industry title*
11	Agriculture, Forestry, Fishing and Hunting
21	Mining
22	Utilities
23	Construction
31 to 33	Manufacturing
42	Wholesale Trade
44 to 45	Retail Trade
48 to 49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate Rental and Leasing
54	Professional, Scientific, and Technical Services
55	Management of Companies and Enterprises
56	Administrative and Support and Waste Management and Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration)

Employee Size Categories	
Category	*Number of Employees*
1	1-9
2	10-49
3	50-249
4	250-999
5	> 1000

RU – Respirator use

P(RU) – Probability of respirator use

i – represents the 2-digit NAICS code, i=1,2,…,19

j – the employee size categories j=1,2,…,5

k – represents a specific entity with cell i,j

N – the total number of entities in the complete sample purchased from D&B.

Ni – the total number of entities in the complete sample for NAICS code i

Nj – the total number of entities in the complete sample for employee size category j

Nij – the total number of entities in the complete sample for NAICS code i and employee size category j

N1 – the total number of entities in the first wave sample from D&B.

N1i – the total number of entities in the first wave sample for NAICS code i

N1j – the total number of entities in the first wave sample for employee size category j

N1ij – the total number of entities in the first wave sample for NAICS code i and employee size category j

N2 – the total number of entities in the second wave sample from D&B.

N2i – the total number of entities in the second wave sample for NAICS code i

N2j – the total number of entities in the second wave sample for employee size category j

N2ij – the total number of entities in the second wave sample for NAICS code i and employee size category j

n - the number of respondents who are members of the target population for a cell or cells represented by upper case N.

r - the number of respondents who are members of the target population that use a respirator of the equivalent cell or cells represented by upper case N

nr - the number of respondents who are members of the target population that do use a respirator of the equivalent cell or cells represented by upper case N

Sample Design and Sample Sizes

The first step will be purchasing sample records from the data provider that are distributed across all targeted industries and that is greater than the number of cases that Abt expects to invite to participate. For this purpose, Abt purchased 760,000 records from D&B.

Sampling will be performed in three waves. The expected sample sizes for each wave are shown in Table 1. For the first wave we will select 300,000 entities from all industry and employee size cells within our

760,000-company sample. In preparation for selecting wave 2, Abt will review responses from Wave 1 to determine which sample cells are underrepresented, and which cells have a high rate of companies using respirators. The sample for Wave 2 will be selected from the remaining unused cases to ensure that underrepresented cells have an increased number of records. There will also be some sample records selected to provide more results in cells with higher respirator use to support detailed analysis of these industries, which requires more participating companies. The sample for Wave 3 will repeat this process to further refine the representation of companies across the entire sample frame.

Table 1. Sample Sizes for Each Wave

Wave	Available Sample (N)	Selected Sample
1	760,000	300,000
2	460,000	150,000
3	310,000	150,000
Total		**600,000**

Tables 2 and 3 assume that all entities in each wave of the sample are members of the target population. Currently, Abt does not know this probability. As a result, the expected number of respondents that are screened and found to be part of the target population will be less than the expected respondent size. To demonstrate the drop in respondent size the assumption is that 90% of the entities in the D&B sample are members of the target population for the first wave. For subsequent waves, the expectation is that 95% of the second wave and 99% of the third wave are part of the target population. The improvement in this percentage is a result of Abt's modeling efforts to identify members of the target population and remove entities from the survey that are not part of the target population.

Table 2. Expected Number of Respondents Assuming a 5% Response Rate and 90% Membership Rate

Wave	Respondents (n)	Members of Target Population
1	15,000	13,500
2	7,500	7,125
3	7,500	7,425
Total	**30,000**	**28,050**

For Table 3 Abt assumes that 4% of entities across all NAICS codes and employee sizes will use respirators. Due to the nature of the adaptive sampling, the unweighted expected percentage of the N entities that respond and are part of the target population is: 4.67%. In addition, the expected percentage of the N entities that respond, use respirators, and are part of the target population is 0.187%.

Table 3. Expected Number of Respondents that Use Respirators and are Members of the Target Population

Wave	Members of Target Population	Respirators (*r*)	No Respirator (*nr*)
1	13,500	540	12,960
2	7,125	285	6,840
3	7,425	297	7,128
Total	**28,050**	**1,122**	**26,928**

NAICS by Size Cells

Currently there are 19 industrial NAICS groups and 5 company employee size categories that have been defined for the SRUP sample frame. The intersection of these NAICS groups and employee size categories creates 95 unique cells. Abt's goal is to obtain sufficient information about respirator use in each of these cells, wherever possible, to support country-level estimates of respirator use. Given that being a member of the target population and the probability of respirator use varies across these cells, it may be difficult to obtain a sufficient sample size in each cell to accurately estimate prevalence for all 95 cells in the study. The number of available entities varies greatly among these 95 cells, with several having less than 100 entities. Given the expected response rate of 4.67% and respirator use rate of 0.187% it is likely that some cells will have few, if any, respondents that use respirators. Low sample size due to response rate will make obtaining a quality estimate of P(RU) challenging. Abt will explore analytic methodologies that allow accurate estimate of prevalence within cells with a low response rate or low respirator usage rate. This will include collapsing cells, modeling probabilities, and small area estimation.

First Wave

Abt will use the available D&B data to determine how many entities to contact from each of the 95 cells. Table 5 shows the sample pull from the D&B data base and includes the count of entities in each of the 95 cells. The value for the number of entities in each of the 95 cells from the sample obtained from D&B is shown in this table. Note that some NAICS codes have been collapsed for the purpose of sampling. These are highlighted in Table 5. There are a small number of entities that report no employees and are shown in the "0" column in Table 5. Abt recommends dropping these entities from the survey as they represent only 11 out of 760,000 entities. Similarly, there are 584 entities that did not have a NAICS code and are shown in the row indicated "Missing." Abt recommends dropping these entities from the survey as they represent less than 1% of the total sample.

Table 5. Values for N; NAICS 2-Digit by Employee Category

	Employee Category						
NAICS	0	1-9	10-49	50-249	250-999	>1000	Total
Missing	0	225	83	239	32	5	584
11	0	21162	916	675	216	34	23003
21	0	8001	2282	1119	335	56	11793
22	0	1927	461	1314	517	146	4365
23	1	57975	5698	6315	1502	154	71645
31	0	24407	3956	3522	1369	223	33477
32	2	4955	1400	6067	2143	390	14957
33	2	5278	1509	10784	4699	832	23104
31-33		34640	6865	20373	8211	1445	71538
42	3	17918	3611	6216	2088	320	30156
44	0	37762	7861	12835	3352	542	62352
45	0	4673	381	4458	1635	154	11301
44-45		42435	8242	17293	4987	696	73653
48	0	28168	1825	3579	1357	254	35183
49	0	3559	338	1023	541	112	5573
48-49		31727	2163	4602	1898	366	40756
51	0	14796	1570	3535	1460	346	21707
52	1	19369	2124	5111	2313	686	29604
53	0	21015	1713	2206	750	118	25802
54	1	46883	6464	9092	3072	612	66124
55	0	3792	383	360	139	61	4735
56	0	25789	2838	5542	2545	725	37439
61	0	11602	3535	14436	3795	630	33998
62	1	49779	10848	16913	7780	2226	87547
71	0	15079	1607	2235	663	130	19714
72	0	15606	21534	10169	2945	535	50789
81	0	46846	3751	3403	924	124	55048
Total	11	486566	86688	131148	46172	9415	760000

Abt will focus in the first sample wave on obtaining information in all 95 cells as the time available for data collection allows, at most, three waves of sample release. The 95 cells are defined using the 2-digit NAICS codes, however, Abt will explore the possibility of drilling down further to 3-digit codes to determine if there are subgroups within an industry that have higher rates of respirator use. This may be utilized when selecting sample records in subsequent waves if information in the initial wave indicates a sub-group has a higher rate of respirator usage.

Abt will stratify the sample using the 95 cells as the primary stratification variable. Within each stratum we will randomly select n_{ij} entities for inclusion in the first sample wave. The sample size in each stratum is shown in the sample size spreadsheet provided in Table 6.

Method for Selecting Sample

Abt proposes sampling 300,000 entities in the first wave. Given that there are 95 cells, any cell with less than 3,158 (300,000/95) will be set equal to the number of entities in the D&B sample pull. All other cells will be increased until a total of 300,000 entities are included in the total sample for the first wave. As the totals are increased, any cell that reaches its maximum will be set equal to that value. The result is that Abt will sample all the entities available in 63 of the 95 cells in the first wave. The remaining 460,000 entities of the D&B sample will be saved for the second and third wave.

Table 6. First Wave Sample Allocation

	Employee Category					
NAICS 2-Digit	1-9	10-49	50-249	250-999	> 1000	Total
11	5999	916	675	216	34	7840
21	5999	2282	1119	335	56	9791
22	1927	461	1314	517	146	4365
23	5999	5698	5999	1502	154	19352
31-33	5999	5999	5999	5999	1445	25441
42	5999	3611	5999	2088	320	18017
44-45	5999	5999	5999	4987	696	23680
48-49	5999	2163	4602	1898	366	15028
51	5999	1570	3535	1460	346	12910
52	5999	2124	5111	2313	686	16233
53	5999	1713	2206	750	118	10786
54	5999	5999	5999	3072	612	21681
55	3792	383	360	139	61	4735
56	5999	2838	5542	2545	725	17649
61	5999	3535	5999	3795	630	19958
62	5999	5999	5999	5999	2226	26222
71	5999	1607	2235	663	130	10634
72	5999	5999	5999	2945	535	21477
81	5999	3751	3403	924	124	14201
Total	107702	62647	78094	42147	9410	300000

Second Wave

The second wave sample specifications will be informed by the analysis of the data collected in the first wave.

Information will be collected from each entity that is sampled in the first wave. This will include determining whether the entity is part of the target population. All entities that are not part of the target population will be removed from the collected data and not used to determine the prevalence of respirator use.

Based on Abt's determination of target population membership, Abt will fit a logistic regression model that estimates the probability of being in the target population conditional on a set of covariates (see Appendix 1). This probability, p_i, will be used to determine if an entity is sampled in the second round and be used for weighting purposes. For each entity a random uniform [0, 1] covariate will be generated which is called x_i. If $x_i <= p_i$ then the entity is available for inclusion in the second sample.

It is expected that 32 of the 95 cells will still have sample remaining after the first wave. Each cell will be given a weight based on the variance of P(RU) and the size of the remaining sample. The process involves the following steps:

First obtain the probability of respirator use for each cell. This is $P(ru)_{ij}=\frac{r_{ij}}{n_{ij}}$. Next, obtain the number of remaining entities from the complete sample for each of the 95 cells. This is $NR_{ij}=N_{ij}-N\,1_{ij}$.

If NR_{ij} is 0, then no further sampling will be conducted in that cell as all N_{ij} entities have been sampled.

The selection weight, w_{ij}, for remaining sample will be:

$$w_{ij}=VAR\left(P(ru)_{ij}\right)*\log\left(NR_{ij}\right)$$

Abt recommends using the LOG function in this equation to spread the sample more evenly to all the remaining cells but still concentrate sample in the cells that contains the largest number of entities.

Next, Abt will recalibrate the selection weights so that they sum to 1.0. This will provide a probability for each of the 32 cells. The actual allocation for each cell for the second wave will be the recalibrate selection weighting probability multiplied by 150,000. If this value is more than the remaining entities for a cell, then all remaining entities will be sampled and the remainder of the sample will be allocated to different cells proportional to the cell selection weights.

It is expected that some of the entities in the second wave sample are not part of the target population. For each entity in *N2*, Abt will use the p_i from the logistic regression model shown in Appendix 1 to determine if that entity is a member of the target population.

Estimating True Population

The expected number of entities available for the second wave in each cell will be the sum of the p_i for any given cell. This will be used as an estimate of the target population members available in the second wave sample. However, the actual number included will be different as this will be randomly determined based on the p_i.

Third Wave

The third wave will be conducted in a manner similar to the second wave. However, Abt may take a more purposive approach in allocation of the sample to certain cells by increasing or decreasing the sample determined by using the selection weights. It is also likely that the entire sample will not be used if there is reason to believe that additional sampling is not required in certain cells.

Abt will consider using specific criteria to stop further sampling for a given cell. Obviously, any cell that has been censused will not be further sampled in the third wave. Another stop criterion Abt will consider is when a cell has low value P(ru). A low probability of use would also result in a low cell weight as small probabilities have small variances.

Possible stop criteria include:

- P(*ru*) < 0.01
- Measurement error of RU less than 0.05.

Extend fielding until June

Abt is extending fielding through June to allow for sufficient time to complete the third wave of sampling while keeping approximately 4-5 weeks between the initial and reminder mailing for each wave.

Appendices

Appendix 1

Probability of being a member of the target population or frame inclusion

Abt will take information about the entities that is available from the information in the D&B database or from information obtained from websites or responses to the survey. From this information Abt will determine what variables can be used as a predictor to determine if that entity is part of the target population and belongs on the frame. This would include types of information such as the existence of:

1) Phone number
2) Website/URL
3) Email
4) Name of contact person
5) Location, region of country
6) NAICS code
7) Employee size
8) Other informative variables

Abt will explore the possibility of using the name of the company for modeling purposes. This might include breaking the name down, looking for common words like INC, LLC, etc. These could then be used as predicators for frame inclusion. There may also be an opportunity to utilize text mining where we have a very large matrix of every word that is used in a name and correlate it with being a member of the target population. For each record, Abt would determine if they are or are not part of the target population. This is a 0/1 binary outcome with 1 meaning that Abt has determined they are part of the target population and 0 otherwise. The information from the list above will serve as predictors.

The next step is to fit a logistic model or use a machine learning/AI technique such as a regression tree. By using predicted value from this model, the probability that each entity is part of the target population can be determined. For the second round of sampling, this probability (call it *pi*) will be used to randomly determine if Abt attempts to contact this entity (part of target population) or does not attempt to contact (not part of target population). This will be accomplished by generating a random uniform variate between 0 and 1 for each entity, x_i. If $x_i <= p_i$ then Abt will contact the entity; otherwise the entity will not be contacted by Abt.

In addition, these probabilities will be used to determine control totals for developing survey weights that will be used to estimate the probability to RU.

Appendix 2

OMB Reconfigured Content:

Research Goal: Determine respirator use among companies that either service Americans or employ American workers.

Target Population: Firms that either employ individuals living in America or service American individuals in-country.

Challenges: Larger companies can have multiple locations that are separate from individual locations or establishments so have a great need to separate 'site' from 'headquarters'. There is a need to establish respirator use estimates across specific industries and across the size (i.e. number of employees) of the individual location or establishment. Additional challenges occur in that it is well known that certain industries would naturally have low respirator use; for example, those that are in technology or other areas where interpersonal exchanges may be rare, no hazardous materials are being handled, and the risk of any infectious disease spread would be expected to be low.

Frame: There is no single comprehensive frame available to use as a sampling frame for this project. The previous NIOSH study used a frame provided by the Bureau of Labor Statistics (BLS). The BLS frame is unavailable and alternate frame options are thought to not provide comprehensive coverage of the target population. An unbiased sample frame should include a listing of all the entities of a target population.

Data Source: Dunn and Bradstreet (D&B) offers a large listing of business entities that are part of the target population, but Abt believes that D&B does not offer complete coverage of the target population. As a result, the likelihood of frame errors and biases exists. Examples of this include having corporate headquarter information but no detailed location information; or a lack of coverage for small businesses; or issues in contacting businesses in certain categories (e.g. small business agriculture). Other similar sources have been explored and all have been assessed to undercover the target population. Abt needs to obtain firm level information and be able to remove establishments in the D&B database. These limitations force the use of an adaptive sampling strategy and a move away from traditional survey assessment metrics.

Sampling Design: Abt will take an adaptive sampling approach for this project. Since there are frame quality issues, Abt will field the data in three waves. Those waves will involve an initial selection and fielding, followed by a response assessment at the intersection of industry and groupings for the number of employees. During each wave of fielding, Abt will assess whether we have duplicate records for multiple locations at the same organization and will explore how many responses are obtained for each of the intersection groups (NAICS by employee size). Armed with that information and with more businesses in the D&B frame, Abt will do outreach in subsequent waves based on obtaining as many unique responses as possible in each of our intersection groups. Abt anticipates having to augment the current frame but are unsure at this time as to the best mechanism to hit our targets, hence the need for an adaptive sample. Options to increase coverage could include:

1. Purchasing frame data from additional sources in targeted industries or by specific company size.
2. Augmenting the frame with data scraped from web-searches that are targeted toward certain intersection groups.

3. Augmenting the outreach with a voluntary response survey option, whose link would be sent to consumer, trade, and other advocacy groups to increase coverage.

All of these options impact the potential for estimation and statistical methodology, however, given the limitations in the D&B sampling frame and the potential for both redundancy and undercoverage, the proposed methodology is the best available at this time for this work.

Nonresponse: Nonresponse is certainly a concern with the current project and has the potential to add additional biases to the potential frame biases for the current work. Without an adequate frame, a traditional response rate calculation is not possible, and we will opt for a participation rate to showcase how many businesses participated based on how many were deemed appropriate for outreach. Given the adaptive sampling options discussed above, some options (such as a voluntary response component) will create additional challenges to estimating a global survey response rate. Abt is targeting for 28,050 completed responses across all industry/business size groupings.

Statistical Power: Abt expects to obtain 28,050 completed responses. Based on the previous NIOSH Survey, there were 19 industries that used respirators with a rate of 3.7% and a margin of error of 0.3%. In the current study, with 28,050 completes Abt anticipates a margin of error of 0.3% for a 95% Wald confidence interval on the population level estimate of respirator use. The target for our smallest intersection group of industry by employee category has 34 entities. We expect 1 complete which would yield a margin of error of 5.7% for a 95% Wald confidence interval on the respirator use in that group.