APPENDIX F

SAMPLE SIZE CALCULATION FOR RANDOM AUDITS

Sample size calculations can be done for the set of mines as a whole or for different sub-groups within the set of mines. ERG has calculated values for six canvass codes used by MSHA: anthracite coal, bituminous coal, sand and gravel, stone, non-metal, and metal. To calculate the sample sizes, ERG used 2011 Part 50 data made available by NIOSH in our calculations.[1]

The first issue we encountered was that the data on injuries in these data are skewed and highly non-normal. ERG used a standard transformation on these data to allow for the data to approach normality; we recalibrated each value of the number of injuries at a mine as $ln(y+1)$ where "*ln*" is the natural log function and y is the number of injuries.[2] This transformation adjusts the data for large outliers (i.e., mines with a large number of injuries) that cause the variance in the sample to be large.[3] ERG performed the sample size calculations using the transformed values. We present the information below transformed back to the number of injuries below. Table F-1 provides relevant values for the sample size calculations from the 2011 Part 50 data.

Table F-1 - Relevant Values for Sample Size Calculations from 2011 Part 50 Data

Category of Data	Canvass Codes					
	Anthracite Coal	Bituminous Coal	Sand and Gravel	Stone	Non-Metal	Metal
Number of mines	242	2,240	6,525	4,283	607	305
Number of injuries	57	3,582	654	1,694	483	1,048
Average	0.2355	1.5991	0.1002	0.3955	0.7957	3.4361
ln(average +1)	0.2115	0.9551	0.0955	0.3333	0.5854	1.4898
Standard deviation of transformed variable	0.3500	0.7986	0.2241	0.4383	0.5971	1.0444

To calculate the sample sizes, ERG used a basic sample size calculation for estimating a mean value from a sample for a continuous variable.[4] The purpose of the calculation was to determine a number of audits necessary to detect a number of underreported injuries as being statistically significant. For example, in bituminous coal, how many audits are needed to determine if there are 500 underreported injuries in that canvass code? Naturally, MSHA does not need to find 500 underreported injuries among the audits (which are a sample). What we are looking for is whether the sample mean number of injuries per audit would indicate whether the total number of injuries among all bituminous coal mines is 500 injuries more than the reported amount. In this case, an estimate of the total number of injuries among the population (based on the audits) is the sample mean (injuries per audit) multiplied by the *total* number (audited and non-audited) of mines in the bituminous coal canvass code.

It was not possible, however, to use a uniform set of values across all canvass codes for the number of underreported injuries. Sample sizes depend on the mean value from the NIOSH data and the

[1] http://www.cdc.gov/niosh/mining/data/default.html. ERG excluded mines that were not in operation during the year from the data we used.
[2] This is one formulation of a Box-Cox transformation, a standard method of normalizing data; see http://robjhyndman.com/hyndsight/transformations/ for more details.
[3] From a sample size calculation perspective, the use of the "raw" data is problematic since most of the observations are less than 10, but a few observations are very large; this leads to a large variance in the population and thus, large sample sizes needed. Additionally, the standard formula for sample size is based on a normal distribution. ERG also investigated the use of alternative distribution assumptions (e.g., Poisson, Weibull), but none of the alternatives were deemed appropriate.
[4] https://www.dssresearch.com/knowledgecenter/toolkitcalculators/samplesizecalculators.aspx.

standard deviation from the NIOSH data. For example, relatively few audits (14) are needed to detect underreporting of 50 injuries in the anthracite coal canvass code dues to the small mean and relatively small standard deviations. However, detecting underreporting of 50 injuries in bituminous coal would require an infeasible number of audits (> 1,000) due to a larger mean and standard deviation. Thus, ERG selected what we felt were reasonable set of injuries to detect as underreporting. In all canvass codes, the smallest 2-3 values we started are most likely infeasible still, but we included them to provide perspective on how sample size varies with accuracy (number of injuries to detect as significant underreporting).

ERG also corrected the estimates for population size by applying the standard "finite population correction" (FPC) to each estimated sample size. The FPC is defined as $n/(1+(n/N))$ where n is the unadjusted value and N is the population size. In the tables below and in the text we report the FPC-adjusted sample size estimates.

Tables F-2 through F-7 provide the results of our calculations. These are discussed in more detailed in the text of the report.

Table F-2 – Number of Audits Needed to Detect Statistically Significant Differences at 95 Percent Confidence, Anthracite Coal Canvass Code

Number of Underreported Injuries to Detect as Statistically Significant	Number of Audits Needed
10	137
20	60
30	32
40	20
50	14
60	10
70	7
80	6

Table F-3 – Number of Audits Needed to Detect Statistically Significant Differences at 95 Percent Confidence, Bituminous Coal Canvass Code

Number of Underreported Injuries to Detect as Statistically Significant	Number of Audits Needed
500	229
600	166
700	127
800	100
900	81
1000	67
1100	57
1200	48

Table F-4 – Number of Audits Needed to Detect Statistically Significant Differences at 95 Percent Confidence, Sand and Gravel Canvass Code

Number of Underreported Injuries to Detect as Statistically Significant	Number of Audits Needed
100	635
200	176
300	81
400	46
500	30
600	21
700	16
800	12

Table F-5 – Number of Audits Needed to Detect Statistically Significant Differences at 95 Percent Confidence, Stone Canvass Code

Number of Underreported Injuries to Detect as Statistically Significant	Number of Audits Needed
100	1310
200	432
300	207
400	121
500	80
600	57
700	42
800	33

Table F-6 – Number of Audits Needed to Detect Statistically Significant Differences at 95 Percent Confidence, Non-Metal Canvass Code

Number of Underreported Injuries to Detect as Statistically Significant	Number of Audits Needed
50	268
100	104
150	53
200	33
250	23
300	16
350	12
400	10

Table F-7 – Number of Audits Needed to Detect Statistically Significant Differences at 95 Percent Confidence, Metal Canvass Code

Number of Underreported Injuries to Detect as Statistically Significant	Number of Audits Needed
100	201
200	103
300	60
400	39
500	28
600	21
700	17
800	10