Supporting Statement B

SUPPORTING STATEMENT FOR

PILOT STUDY AND PROSPECTIVE ANALYSIS OF THE DRAFT REVISED FORM 33,

SAFETY AND HEALTH PROGRAM ASSESMENT WORKSHEET (March 2025)

COLLECTIONS OF INFORMATION EMPLOYING STATISTICAL METHODS

If your collection does not employ statistical methods, just say that and delete the following five questions from the format – UNLESS your proposed information collection is a survey. OMB has recently clarified that Part B must be completed for all survey requests, whether or not statistical analysis will be applied. In addition to statistical analysis, Part B addresses the description of the target group of respondents, the sampling plan, and plans to maximize response rates and address non-response.

When statistical methods will be used to tabulate data, the following documentation must be provided to the extent that it applies to the methods proposed.

Note to the Reviewer

This Information Collection Request (ICR) is for an extension of 3 years to the approved "Pilot Study and Prospective Analysis of the Draft Revised Form 33 (DRF33) Safety and Health Program Assessment Worksheet," which commenced in July 2022

The data analyses plan described in this Supporting Statement B has not changed. Responses below incorporate updates that reflect the remaining data that must be completed to complete the Pilot Study and address the remaining tasks.

1. Describe (including a numerical estimate) the potential respondent universe and any sampling or other respondent selection methods to be used. Data on the number of entities (e.g., establishments, State and local government units, households, or persons) in the universe covered by the collection and in the corresponding sample are to be provided in tabular form for the universe as a whole and for each of the strata in the proposed sample. Indicate expected response rates for the collection as a whole. If the collection had been conducted previously, include the actual response rate achieved during the last collection.

Note: response rate means: Of those in your respondent sample, from what percentage do you expect to get the required information (if this is not a mandatory collection). The non-respondents would include those you could not contact and those you contacted but refused to give the information.

Potential Respondent Universe

The OSHA On-Site Consultation Program [1]is administered by the Agency's Directorate of Cooperative and State Programs (DCSP), Office of Small Business Assistance (OSBA). OSHA provides guidance to state[2] On-site consultation programs through the OSHA Consultation Policies and Procedures Manual (CPPM), CSP 02-00-005, September 29, 2023. The CPPM complies with the requirements of 29 CFR 1908 to specify the framework for administering and managing the OSHA On-Site Consultation Program and establish policies and procedures.

29 CFR Part 1908, Consultation Agreements, contains requirements for Cooperative Agreements between states and the federal Occupational Safety and Health Administration (OSHA) under sections 21(c) of the Occupational Safety and Health Act of 1970 (the Act; 29 U.S.C. 651 et seq.) and section 21(d), the OSHA Compliance Assistance Authorization Act of 1998 (which amends the Act), under which OSHA will utilize state personnel (consultants) to provide consultative services to small- and medium-sized employers or businesses, through the On-Site Consultation programs established by U.S. states and territories. The On-Site Consultation Program is designed to provide small- and medium-sized businesses with occupational safety and health services. OSHA encourages larger sized businesses to implement effective Safety and Health Management Programs (SHMPs) through the Voluntary Protection Programs.

The Consultation Annual Program Plan (CAPP) is a requirement of the annual Cooperative Agreement between OSHA and each state On-Site Consultation program. The CAPP must include the projected number of the different types of consultation services that the Consultation program intends to conduct during the fiscal year ((FY); e.g., consultation visits, compliance assistance activities, training). The activities conducted by state Consultation programs for the Pilot Study, i.e., the pre-test, consultation visits during which comprehensive safety and health program assessments are conducted, Prospective Analysis, data analyses, and follow-up study, will count towards meeting the projected number of consultative services in the CAPP.

As required by the Cooperative Agreement between OSHA and the states, each state Consultation program will maintain at least two consultants in the safety discipline and two consultants in the health discipline, to provide no-cost consultative services to small- and medium-sized businesses nationwide (i.e., 50 states and U.S. territories). A minimum of 200 consultants (i.e., respondents) from state On-Site Consultation programs nationwide are estimated to participate in the Pilot Study. Consultants participating in the studies are from the general pool of trained consultants.

Employers can voluntarily contact their state On-Site Consultation program to request no-cost consultative services provided by consultants to assist with improving occupational safety and health

1

2

conditions. On-Site Consultation programs located within state agencies or universities employ consultants to provide consultative services to employers, off-site or on-site at their workplaces. These include consultation visits at workplaces that involve hazard assessments, safety and health program (SHP) assessments, and/or training.

Consultation visits for the Pilot Study are conducted by consultants at workplaces where employers voluntarily request such services from state On-Site Consultation programs. Consultation programs nationwide typically complete more than 20,000 consultation visits each FY (except for FY 2020 due to the impact of the COVID-19 pandemic). Therefore, **the potential respondent[3] universe of employers for the Pilot Study** comprise small- and medium-sized businesses that voluntarily contact state Consultation programs nationwide to request consultation visits.

Consultation programs' participation in these studies is voluntary. DCSP/OSBA discussed participation in the Pilot Study with Consultation Program Managers (CPMs), who are responsible for overseeing the daily operation of state On-Site Consultation programs.

The approved OMB studies of the DRF33 (i.e., OMB #1218-0280), include the following: a pre-test using a maximum of 20 consultation visits to assess Pilot Study procedures and the information technology platform, and to correct any issues before launching the Pilot Study; 300 Pilot Study consultation visits to small- and medium-sized business workplaces, during which 350 SHP assessments will be conducted by consultants using the DRF33; a Prospective Analysis involving injury and illness data collection by consultants 12 months after the consultation visits to assess any impact of the application of the DRF33; data analyses to determine validity and reliability; and a follow-up study involving a maximum of 30 consultation visits, when necessary to assess any updates to the DRF33 resulting from data analyses and other findings.

Following OMB approval, OSHA commenced the Pilot Study of the DRF33 in July 2022 with a pre-test; and consultants have completed 280 pilot consultation visits to workplaces using the DRF33 to do 326 comprehensive SHP assessments out of the 350 planned.

OSHA is requesting additional time to complete the remaining studies:

- 20 pilot consultation visits with 24 SHP assessments;
- Prospective Injury and Illness Data Analysis: All 300 pilot consultation visits require prospective injury and illness data collected from employers 12 months from the visit date by consultants, via phone, email, fax, etcetera;
- Data analyses to assess validity and reliability of the DRF33; and
- Any follow-up consultation visits with SHP assessments required.

<u>Pilot Study</u>

[3] Consultants are referred to as respondents in Table 1 for data collection purpose, because consultants will be responsible for collecting SHP information when they assess employers' workplaces, and will complete the DRF33 with their findings and recommendations. Employers will not complete the DRF33 but will receive a copy of the completed DRF33 for their workplace and can use the information to improve their SHPs.

Since July 2022, DCSP/OSBA has conducted the approved pilot study in collaboration with subject matter experts (e.g., occupational safety and health experts, psychometricians/statisticians, and economists) from OSHA and state On-Site Consultation programs operated by the University of South Florida and the Georgia Institute of Technology. The execution of the Pilot Study has also been coordinated with the National Association of Occupational Safety and Health Consultation Programs (OSHCON), made up of CPMs nationwide.

The Pilot Study has involved about 200 trained safety and health consultants from state On-Site Consultation programs nationwide using the DRF33 (instead of the current OSHA Form 33) to conduct comprehensive SHP assessments of small- and medium-sized workplaces during consultation visits when employers request such consultative services (provided to them at no cost).

The DRF33 was uploaded into Qualtrics, a cloud-based data entry tool, for consultants to enter SHP assessment information (i.e., a summary of findings based upon workplace SHP assessments, ratings for each of the 52 attributes (i.e., SHP assessment criteria), and recommendations to the employer. The DRF33 is supported by a draft revised ARG that contains detailed descriptions of all attributes, updated guidance for assessing the implementation of each attribute at a workplace, and several work aids for consultants to use. The ARG was also uploaded into Qualtrics for easy access and as a job aid for consultants. Consultants participating in the Pilot Study follow the policies and procedures in the CPPM for conducting consultation visits, including SHP assessments.

There are seven (7) core elements of the DRF33 (i.e., Management Leadership; Worker Participation; Hazard Identification and Assessment; Hazard Prevention and Control; Education and Training; Program Evaluation and Improvement; and Communication and Coordination for Host Employers, Contractors and Staffing Agencies (i.e., the multi-employer element)). Each core element is described by a number of attributes. The total number of attributes in the DRF33 is fifty-two (52), and four of these attributes directly assess the multi-employer element.

A comprehensive SHP assessment means that consultants (i.e., respondents) will assess the level of implementation of each of the 52 attributes at a workplace and assign ratings recorded in the DRF33. The possible ratings for an attribute range from 0 to 3. In the case of a workplace where there is only one employer at all times and all workers in the workplace work directly for this employer, the multi-employer element would not apply, and consequently, consultants will not be able to assess the implementation of the multi-employer attributes but must assess the remaining 48 attributes. Therefore, it is important to emphasize that consultants must evaluate every workplace assessed in these studies with regard to all the 48 DRF33 attributes, which do not explicitly pertain to the multi-employer core element. Additionally, the four multi-employer attributes must be rated at every workplace that may have other workers present at any time (e.g., temporary, seasonal, and contractual workers) in addition to the host employer's workers.

Core element scores are derived by summing the ratings for attributes in each core element and subsequently placing these summed scores onto a 0 to 100-point scale. A weighted total score is derived analogously across all attributes (except the four multi-employer attributes). An unweighted total score is produced by simply averaging all core elements (except the multi-employer element). The multi-employer element will be excluded from the total scores due to its systematic missing data, but this element will be included in the primary and secondary analyses of multi-employer worksites whenever it is appropriate. Consultants will also input a summary of SHP assessment findings and recommendations for employers' workplaces into the DRF33. In addition, consultants may choose to provide feedback on the application of the DRF33 and ARG through six optional survey questions. The data collected by consultants in the DRF33 will be used to conduct validity and reliability analyses to determine the effectiveness of the DRF33 in assessing workplace SHPs.

TRC and DART Calculations

Although the total and subscale DRF33 scores will serve as the primary measures in the analyses that will be conducted, measures regarding other workplace level covariates will also be obtained. Since the overall goal of SHPs is to prevent the occurrence of illnesses and injuries that may result from exposure to hazardous workplace conditions and hazardous work practices, measures of the total recordable case rate (TRC), and the days away, restricted or transferred rate (DART) will be used.

Since DART and TRC are indices of occupational illness and injury, one would hypothesize that these rates would be lower for workplaces with strong or effective SHPs, rather than those with weak or ineffective programs. Indeed, Shea, De Cieri, Donohue, Cooper, and Sheehan (2016) suggested that some SHP activities that are leading indicators[4] are negatively related to outcomes like workplace injury or compensation claims. Autenrieth, Brazile, Sandfort, Douphrate Roman-Muniz, and Reynolds (2016) reached similar conclusions when examining TRC and DART with the OSHA Form 33 that is currently in use. However, their conclusions generally hold only when all attributes of the worksheet have been assessed and rated by consultants.

The TRC and DART used in these studies will be calculated as follows:

- H = logged cases resulting in days away from work
- I = logged cases resulting in job transfer or restriction
- J = other logged cases
- E = estimated hours worked by all employees at the workplace during the 12 months preceding (i.e., retrospective analysis) or following (i.e., prospective analysis) the Pilot Study consultation visit
- S = a scale factor equal to 200,000 (equivalent to 100 full-time employees working 40 hours per week for 50 weeks per year).

Then:

[4] "Leading indicators are measures intended to predict the occurrence of events in the future. Leading indicators are proactive, preventative and predictive measures that provide information about the effective performance of safety and health program activities that can drive the control of workplace hazards. While lagging indicators are measures of the occurrence and frequency of events in the past such as the number and rates of injuries, illnesses, and fatalities" (- OSHA's Recommended Practices for Safety and Health Programs).

- TRC = [(H + I + J) / E] * S
- DART = [(H + I) / E] *S.

Notes:

- Values for H, I, and J will be acquired from OSHA's Form 300, Log of Work-Related Injuries and Illnesses, maintained at the workplace.
- The injury and illness data in OSHA's Form 300 is summarized in Form 300A, Summary of Work Related Injuries and Illnesses.

The TRC gives the total cases one would expect from 100 full-time employees at a given workplace, whereas DART yields a similar expectation with respect to only cases resulting in days away from work, job transfer, or job restriction. These indices are scaled to make workplaces of various sizes comparable.

Both TRC and DART must be calculated for a given period of time. Care must be taken to make the time period long enough to register adverse events which may be infrequent at a given workplace yet short enough to provide fidelity to the outcomes at that workplace so as not to average them out. This study will base calculations of TRC and DART on the 12 months preceding (retrospective) and following (prospective) the Pilot Study consultation visits. Therefore, the term "retrospective" might be considered a "baseline" or "current" measure of TRC or DART. The term "retrospective" is used because those calculations are based on logged observations from a period of time preceding the Pilot Study consultation visits even though they represent the most current state of affairs at that time.

The data analyses described in #2 of this document will be performed after the Pilot Study consultation visits are completed using the DRF33 and retrospective covariate measures. Prospective Analysis involving prospective estimates of TRC and DART will occur roughly 12 months after the Pilot Study consultation visits.

During the Pilot Study consultation visits, consultants collect copies of injury and illness Forms 300 and 300A for the calendar year preceding the consultation visit and Form 300 for the year of the consultation visit (up to the date of the visit). Consultants also collect information about the number of employees at the workplace and the number of hours worked by all employees, for the year of the consultation visit – up to the date of the visit. Consultants email all collected injury and illness data to be uploaded into a secure Dropbox for analysis. Statisticians use the data collected to calculate the retrospective TRC and DART for the 12-month period prior to the Pilot Study consultation visits. Similar data will be collected by consultants for the 12-month period following the Pilot Study (from workplaces that previously received Pilot Study consultation visits); for statisticians to calculate the prospective TRC and DART.

Design

OSHA developed the DRF33 for state On-Site Consultation programs that provide no-cost consultative services to small- and medium-sized businesses nationwide (i.e., similar to how the current OSHA Form 33 is applied). Therefore, it is essential to assess the effectiveness of the DRF33 as a tool for measuring the SHPs of small- and medium-sized businesses. State On-Site Consultation programs were geographically grouped into ten (10) regions of the nation. The five (5) industry sectors selected for the Pilot Study represent the top 5 industries most visited by On-Site Consultation programs nationwide.

Information gathered from expert consultants suggested that workplace size and type of industry evaluated would likely influence scores on the DRF33 more than other external variables. Additionally, there is a strong desire to include consultation visits from all of the ten (10) regions of the nation. Therefore, these variables will be used as stratification variables in a convenience sampling plan.

The Pilot Study to assess the reliability and validity of the DRF33 is based on a 2 (workplace size) x 5 (industry sector) x 10 (region) factorial design. For every region, an equal number of small- (less than or equal to 25 employees on-site) and medium-sized (greater than 25 but fewer than 250 employees on-site) workplaces will be sampled from each of five pre-specified industry sectors. The five industry sectors were chosen using data from the OSHA Information System (OIS) for at least three years (e.g., FY2018-FY2020). Specifically, the most frequently surveyed sectors (i.e., North American Industry Classification System (NAICS) code) were identified using data from across the nation. From this subset of industry sectors, five sectors with sufficient representation across all 10 regions were selected for the study. Each selected sector should possess the requisite number of small- and medium-sized workplaces in each region.

Sample characteristics

The Pilot Study will use a stratified convenience sample with workplace size, industry sector, and region as stratification variables. As mentioned above, this will yield a 2 x 5 x 10 = 100 cell design. There will be three workplaces sampled in each of the 100 cells of the factorial design for a total sample size of 300 workplaces. As discussed later in this document, the attribute ratings from the DRF33 will be analyzed with an exploratory factor analysis procedure, and it is this procedure that determines the study sample size. A sample size of 300 workplaces leads to a cases per attribute ratio of 5.8 which is in the range of recommendations suggested in the exploratory factor analysis literature (Cattell, 1978; Gorsuch, 1983). Workplaces will be selected by the Consultation Program Managers who oversee state On-Site Consultation programs (using the criteria specified for the study) to represent the most-frequently visited industries in the country with two classifications of company size (i.e., small and medium).

Pilot Study consultation visits are assigned in an effort to meet the study criteria based on the open pool of requests within each Consultation program. The workplaces selected for study visits must not have an active OSHA or State Plan enforcement inspection or involvement at the time of the consultation visit and must be in business for at least 2 years. The employer must not have an employee population of more than 250, corporate wide.

All study measures for the 48 attributes are obtained from every workplace; and the additional four attributes of the multi-employer core element are designed for only multi-employer workplaces.

As mentioned earlier, workplaces request consultation visits from the state On-Site Consultation programs. For this reason, pilot study drop-out is expected to be quite low and would chiefly result from the consultant's inability to conduct an assessment using all attributes from the DRF33 or workplace closure like that previously experienced during large COVID-19 outbreaks. Neither of these conditions directly relate to the quality of the SHP at a given workplace, so the likelihood of informative missing data seems low. Moreover, when any missing data are encountered, a substitute workplace is identified for the stratum in question and a consultant visits the substitute workplace instead. It is expected that this procedure will yield complete data from 300 workplaces.

The Pilot Study involves a comprehensive SHP assessment of 300 workplaces, each requiring five (5) hours to complete. In addition, to evaluate the DRF33 interrater reliability – fifty (50) of the 300 workplaces assessed (i.e., five from each of the 10 regions) will be visited by two (2) consultants. Both consultants will fill out the DRF33 independently based on the information available at the workplace, therefore, the 300 Pilot Study workplace consultation visits will result in 350 SHP assessments. So far, 326 pilot workplace SHP assessments have been conducted by Consultation programs nationwide out of the total number of 350 planned.

<u>Prospective Analysis</u>

Prospective estimates of the TRC and DART will be based on the 12 months following the Pilot Study consultation visits, as described above (i.e., TRC and DART Calculations). These covariates will be analyzed with the same procedures used for the retrospective analyses, although these analyses will take place one year after the Pilot Study consultation visits when prospective covariates are available. The Prospective Analysis does not require any additional consultation visits, therefore, there will be no SHP assessments using the DRF33. Consultants will collect injury and illness data from employers via phone, fax, email, etcetera.

Pre-Test and Follow-Up

All pre-test SHP assessments for the Pilot Study were conducted to ensure that all procedures and information technology platforms will function as designed during the Pilot Study. Any failings observed in procedures or technology used during the pre-test was fixed before launching the Pilot Study. The pre-test was conducted in only one of the ten regions with fewer than the maximum of twenty (20) workplace SHP assessments estimated.

Similarly, a follow-up will be conducted, using a maximum of thirty (30) workplace SHP assessments, only if data analyses and other findings result in updates to the DRF33 and ARG that require verification of effectiveness before implementing operationally.

Table 1: Summary of Data Collection Plan

Activity	Type of Respondents[5]	No. of Workplaces Visited	No. of Respondents[6]	No. of Responses per Respondent	Total No. of Responses	Average Burden (Hours)	Total Burden (Hours)
Pilot SHP Assessments	Consultants	20	20	1.2	24[7]	5	120
Follow-up SHP Assessments	Consultants	30	20	1.5	30	5	150
Total		**50**			**54**	**5**	**270**
Prospective Analysis[8]	Consultants	0	200	1.5	**300**	0.5	**150**

2. Describe the procedures for the collection of information including:

- **Statistical methodology for stratification and sample selection**,
- **Estimation procedure,**
- **Degree of accuracy needed for the purpose described in the justification,**
- **Unusual problems requiring specialized sampling procedures, and**
- **Any use of periodic (less frequent than annual) data collection cycles to reduce burden.**
- **If you are selecting a uniform respondent universe, you may be using simply a random numbers table to select a sample.**

SHP Assessment Data Analyses and Prospective Analysis

Consultants use the web-based DRF33 to collect information on comprehensive SHP assessments of workplaces that they conduct when employers request such services. DRF33 is transcribed into a web-based data entry tool (e.g., Qualtrics) into which consultants enter their evaluations (e.g., summary of

[5] The Agency categorizes respondents reflected in Table 1 as "State, Local and Tribal Governments" (i.e., consultants that will complete the DRF33 using information collected from employers' workplaces).

[6] OSHA will engage a minimum of 200 consultants to participate in the Prospective Analysis; consultants participating in the study will be selected from a pool of trained consultants.

[7] Originally, OSHA estimated engaging 200 consultants to conduct 300 pilot workplace consultation visits (i.e., an estimate of 20 consultants from each of the 10 regions), that would result in 350 SHP assessments; consultants participating in the studies were selected from a pool of trained consultants. OSHA estimates engaging an average of 20 consultants to conduct the remaining pilot consultation visits.

[8] The Prospective Analysis will not involve any additional Pilot Study consultation visits, therefore, there will be no SHP assessments using the DRF33. Consultants will collect injury and illness data from employers via phone, fax, email, etcetera.

findings based upon workplace SHP assessments, ratings for each of the 48 or 52 attributes (as applicable), and recommendations to the employer).

The DRF33 is administered and scored by core element (i.e., a score is the summation of all ratings for all attributes within an element, or all elements combined). There are seven (7) core elements of the DRF33: Management Leadership; Worker Participation; Hazard Identification and Assessment; Hazard Prevention and Control; Education and Training; Program Evaluation and Improvement; and Communication and Coordination for Host Employers, Contractors and Staffing Agencies (i.e., the multi-employer element).

Attribute ratings (which range from 0 to 3) on each core element are summed and each summed score is transformed into a new score metric ranging from 0 to 100 as follows:

New Core Element Score = 100*Sum Score / (3*ICE)

$$New\ Core\ Element\ Score = \frac{100 * \sum Score}{(3 * ICE)}$$

where ICE denotes the number of DRF33 items associated with a particular core element. Thus, the new core element score (i.e., subscale) would range from 0 to 100 regardless of the number of items which represent that core element. This type of rescaling facilitates the comparison of core element profiles across workplaces. In addition to calculating scores for the core elements, two total DRF33 scores are also produced. First, an unweighted total score is derived from the average of the core element scores, excluding the multi-employer element due to its systematic missing values. Because each of these subscales has a possible range of 0-100, regardless of the number of attributes associated with any core element, the average of these subscale scores is an unweighted mean across the core elements and has an identical range. A second total score is produced. The average rating across each of the 48 attributes associated with core elements other than the multi-employer element is rescaled to range from 0-100. This is accomplished using the following equation:

$$Rescaled\ Weighted\ Average = 100 * \left(\frac{\sum_{k=1}^{48} r_k}{144} \right)$$

Where r_k is the consultant rating (i.e., 0 to 3) for the kth attribute on the DRF33. Because the maximum summated score on these attributes is 3*48 = 144, this formula converts the total summated score into a proportion and then rescales this result to produce a 0-100 range. In this sense, the weighted total score will have the same range as any other DRF33 measure but will be weighted more by core elements with the largest number of attributes and weighted less by those with the fewest number.

It is important to emphasize that consultants must evaluate every workplace assessed in these studies with regard to all of the 48 DRF33 attributes, which do not explicitly pertain to the multi-employer core element. Also, the four multi-employer attributes must be rated at every workplace with other workers (e.g., temporary, seasonal, and contractual workers) in addition to the host employer's workers. Previous research has illustrated that missing attribute data can severely alter the correlations among scores on the (current) Form 33 (OMB # 1218-0110) core elements with external criteria (Autenrieth et al., 2015). This should be avoided. Similarly, a larger sample would be required if the psychometric analyses outlined later in this proposal were based on DRF33 responses containing systematically missing data in which consultants rated only a subset of attributes on the DRF33. In this case, the proposed analyses would be conducted on only complete cases (for all but the multi-employer core element), and number of workplace consultation visits would need to increase to achieve a desired sample size of 300 completed assessments.

Although the total and subscale (i.e., core element) DRF33 scores serve as the primary measures in this investigation, measures regarding other workplace-level covariates will also be obtained. In particular, measures of the total recordable case rate (TRC) and days away, restricted or transferred rate (DART) will be used. Retrospective estimates and prospective estimates of these covariates will be obtained. Retrospective estimates of TRC and DART will be based on the 12-months prior to the Pilot Study consultation visit, whereas, prospective estimates will be acquired 12-months after the visit. The statistical analyses described in this document will be performed after the Pilot Study workplace consultation visits are completed and the DRF33 and retrospective covariate measures have been uploaded into the Qualtrics database. They will be repeated one year after the Pilot Study consultation visits, once prospective covariates are available from the 300 workplaces that were previously assessed.

All consultants' ratings and covariate data gathered at the workplace consultation visit will be entered into a data entry tool built from the Qualtrics platform. The data entry tool will prompt the respondent (consultant) to enter the data associated with each attribute. The tool will offer real-time accuracy checks and secure cloud-based data storage that can be retrieved from multiple sites immediately after input. Similarly, it can be programmed to ensure that all applicable attributes are answered by the respondent. It will operate with common digital devices (e.g., tablet, PC) that has internet capability.

Research Questions for the Pilot Study

1. **Reliability of DRF33 Scores from the On-Site Consultation programs**

Reliability is an important feature of any measurement tool used in an evaluation situation. Moreover, one might examine several types of reliability (Crocker & Algina, 1986; Nunnally & Berstein, 1994). This study proposes to address two widely recognized types of reliability estimation. The first is internal consistency estimation which gets at the cohesiveness of item responses underlying a given score. Ideally, the item responses should covary a substantial amount. The second type of reliability examined in this study is interrater reliability. When raters evaluate workplaces, an important assumption is that rater scores covary to a substantial degree. Interrater reliability estimation indexes the extent to which this occurs. These two types of reliability estimation underlie the following research questions:

a) What is the internal consistency estimate of reliability for total scores based on the 48 attributes from the DRF33?

b) What is the internal consistency estimate of reliability for core element (i.e., subscale) scores for each of the seven core elements?

c) What is the interrater reliability estimate for total scores based on the 48 attributes from the DRF33?

d) What is the interrater reliability estimate for core element (i.e., subscale) scores for each of the seven core elements?

2. **Validity of the DRF33 Scores from the On-Site Consultation programs**

The validity of scores from a measurement instrument is another primary feature that should be demonstrated prior to using those scores in an evaluation context (Crocker & Algina, 1986; Nunnally & Berstein, 1994). Several types of evidence can be assembled to demonstrate the validity of test scores. This study will examine three types.

First, we will acquire evidence on the DRF33's construct validity. Construct validity is a common form of validity evidence for test scores, where hypothesized relationships between test scores and other external variables or constructs are examined.

Second, we intend to examine the correlational structure of attribute ratings from the DRF33 in order to determine how many latent dimensions are measured by the instrument and what those dimensions might be. This is also a form of construct validity, although it is internal to the test scores themselves, and it will be studied with exploratory factor analysis.

Third, we will assess the criterion-related evidence for validity of DFR33 scores. Criterion-related validity examines the relationships between test scores and external performance criteria which are of fundamental interest to the researcher. These criteria may represent the performance status when the test was administered (concurrent validity) or later (predictive validity).

These different types of validity evidence form the basis of the following research questions:

a) How do the DRF33 total and subscale scores differentiate among workplaces of various sizes and in alternative industries that are common across the 10 regions?

b) How many latent factors underlie the DRF33 responses and what is a reasonable interpretation of these factors?

c) How do DRF33 total and subscale (i.e., core element) scores relate to fundamental safety and health outcomes like total recordable case rate (TRC) and days away, restricted or transferred rate (DART)?

<u>Statistical Analyses</u>

A series of univariate analysis of variance (ANOVA) models will be examined in which DRF33 core element scores and both total scores serve as a single dependent measure in nine successive analyses. The ANOVA will assess the statistical significance of mean differences on a given dependent measure when exploring the main effects of industry sector, workplace size, and region. This will answer questions such as “Are the differences between means across the regions so large in the context of sampling error that they suggest that nonzero mean differences actually exist in the population?” Similar questions pertaining to sector and workplace size will be answered. The two-way interactions for sector x workplace size, sector x region, and workplace size x region will be examined in a similar fashion, as will the three-way interaction involving sector x workplace size x region. The familywise Type I error rate associated with the test of each of these seven effects in the ANOVA model will be set to the traditional value of 0.05. When the omnibus test of any effect suggests that population group mean differences exist, then appropriate post hoc tests (e.g., Tukey's Honestly Significant Difference (1949), Scheffe F-test (1959)) will be conducted as a follow-up testing procedure. The sample size for these ANOVA models will be limited to N=300. In other words, the second DRF33 rating performed in 50 of the 300 sampled workplaces will not be used in these analyses (as they would violate the independence of scores assumption).

In addition to the univariate ANOVA models described above, a multivariate ANOVA (MANOVA) model will be performed using scores from each DRF33 core element (other than the multi-employer element) as multiple dependent variables. (Only complete cases are allowed in a MANOVA, and the multi-employer element cannot be evaluated for all workplaces.) This analysis will offer additional information about the group mean differences for each of the seven effects (i.e., main effects and interactions). In particular, this analysis will explore which dimension(s) in the space of the dependent variables will differentiate groups the most. For example, linear combinations involving particular core elements may differentiate industry sectors most, whereas linear combinations of other core elements might distinguish workplaces of various sizes most. The univariate and multivariate ANOVA models described in this section correspond to Research Question 2a and will provide evidence to support the construct validity of the DRF33.

Psychometric Analyses

Reliability Analyses. Two types of reliability analysis will be performed. The first is an analysis of the internal consistency of attribute ratings for a given core element. This will be performed using Cronbach's (1951) alpha coefficient. Cronbach's alpha will be calculated separately for the attributes of each core element and, thus, will yield estimates of reliability for each set of core element scores (i.e., subscales). If subsequent (factor) analysis suggest that some attributes are better aligned with other core elements, then those attributes will be reassigned accordingly, and the analysis of internal consistency will be repeated using the new attribute subsets. For each of these two applications of Cronbach's alpha, attributes that contribute little to the internal consistency of the scale or even decrease it will be identified as possible candidates for elimination from the DRF33. These internal consistency analyses correspond to Research Questions 1a and 1b and will provide one source of evidence about the reliability of DRF33 scores.

The second type of reliability information that will be evaluated is interrater reliability. Fifty of the 300 workplaces sampled in this study will be visited by two consultants who both fill out the DRF33 independently based on the information available at the worksite. These 50 pairs of DRF33 scores will be evaluated with respect to both the consistency and the agreement of scores from each pair of consultants. This assessment will be done separately for each core element as well as the two total scores. Investigation of interrater reliability between the 50 pairs of data points will be conducted using intraclass correlation coefficients (Fleiss, 1981; Shrout & Fleiss, 1979), and will be supplemented with Pearson and Spearman correlation coefficients. These interrater reliability analyses address Research Questions 1c and 1d.

Validity Analyses

Additional evidence supporting the validity of the DRF33 will be obtained from two primary methodological strategies. The first is through factor analysis techniques. Factor analysis attempts to explain the common variance among a set of variables by postulating a set of underlying constructs (factors). In the current context, these constructs are presumably responsible for the attribute ratings on the DRF33, and thus, they can theoretically account for correlations among pairs of attribute ratings. Factor analysis will be used to identify the number and nature of the underlying constructs required to explain the correlation among those ratings. Because standard Pearson correlations between discrete (rather than continuous) responses may underestimate true linear relationships and result in spurious factors, polychoric correlations among pairs of attribute ratings will be used in all factor analyses conducted in this study (Gorsuch, 1983; Lee, Poon & Bentler, 1995). Polychoric correlations, are in essence, estimates of Pearson correlations that would have emerged if the observed discrete responses resulted from categorizing the domain of a normal distribution into successive, discrete intervals (i.e, graded categories like 0, 1, 2, and 3).

Three factor analysis models will be fit to the DRF33 responses. The first will be a traditional exploratory factor analysis (EFA) model in which each of the attributes is allowed to correlate with each latent factor estimated in the model. Specifically, if we presume that each latent factor represents a particular core element (e.g., management leadership, hazard identification, and assessment, etc.), then each attribute will be allowed to correlate with each underlying core element. Ideally, a particular attribute would correlate only with the core element that it was designed to measure, but that is an empirical question that the analysis will address. EFA, by definition, produces orthogonal (i.e. linearly independent) factors). However, there is no reason to believe that the constructs underlying each core element are uncorrelated. For that reason, EFA results will be rotated to allow for correlations among core elements. This will be accomplished using a Promax rotation (Hendrickson & White, 1964).

In addition to the EFA, two other confirmatory factor analysis models will be estimated. The first of these will be a simple structure model (Joreskog, 1966) in which only those attributes designed to measure a given core element will be allowed to correlate with the underlying factor associated with that core element. However, the underlying factors will be allowed to correlate with each other. The final factor analysis model to be explored, referred to as a bifactor model (Gibbons & Hedeker, 1992), will allow each DRF33 attribute to correlate with a general factor that could be interpreted a general safety and health program quality factor. In addition to this correlation with the general factor, each attribute will be allowed to correlate with a second factor that represents that part of the underlying core element unique from the general factor. Any correlations among attributes from different core elements are assumed to result from the general factor.

The fit of each of these aforementioned factor analysis models will be assessed using a variety of indices along with manual inspection of residuals. The first fit index that will be calculated is the root mean square error of approximation (RMSEA; Steiger & Lind, 1980). This index provides both a relative and absolute index of model fit. Lower values of the RMSEA are preferred. RMSEA values less than .08 are deemed reasonable whereas those less than .06 are interpreted as good (Hu & Bentler, 1998). Two information criteria will also be employed to explore relative model fit. These are Akaike's (1973) information criterion (AIC) and the Bayesian information criterion (BIC; Schwarz, 1978). Both indices evaluate misfit of the model with larger values reflecting more misfit. Additionally, both indices incorporate a penalty for more complicated models. Models with greater numbers of estimated parameters are penalized more than those with fewer estimated parameters. The difference in the two criteria is that this penalty for the number of estimated parameters is more severe for the BIC than the AIC. In this sense, the BIC weights parsimony more than the AIC when used to evaluate alternative models.

The three aforementioned factor analysis models will be repeatedly estimated using alternative data sets. First, the data for the 48 attributes that represent core elements other than the multi-employer element will be examined. These data will be available (without missing values) for all 300 workplaces participating in the Pilot Study. Second, the factor analyses described above will be performed again using all 52 DRF33 attributes for only the workplaces with other types of workers (e.g., temporary, contractual, and seasonal workers) in addition to the host employer's workers. Third, the three-factor analyses will be repeated again, but this time using a full correlation matrix built from pairwise correlations that only require complete data for elements in the pair of attributes involved in the correlation (rather than all of the attributes). This will enable us to examine all attributes and all workplaces simultaneously, albeit with a non-Gramian correlation matrix.

Each factor analysis model above will help better understand the correlational structure of attribute ratings on the DRF33. Moreover, when taken together, they will provide substantial information about the internal (construct) validity of the DRF33 as specified in Research Question 2b.

Evidence for the validity of DRF33 core element scores and total scores will also be developed from correlations between these scores and external outcome covariates which should logically covary with the quality of a safety and health program. Specifically, each DRF33 score will be correlated with both the retrospective and prospective measures of TRC and DART using both Pearson and Spearman (rank order) correlation indices. Additionally, each of these covariates will serve as a dependent measure in a multiple regression model in which all DRF33 core element scores (other than the multi-employer element) serve as predictors. This will reveal the unique contribution of each DRF33 core element score to the prediction of the respective covariates as opposed to simply the total contribution assessed using simple correlations. The statistically significant predictors in this model will be identified, and the proportion of variance in the dependent variable will be quantified. A multivariate multiple regression (Finn, 1974) will also be performed to determine if there is a dominant dimension underlying the two covariates that can be explained in a statistically significant fashion by any of the core element scores. The analyses above will be repeated using only the subset of workplaces with ratings for all DRF33 attributes, including those for multi-employer workplaces. The univariate and multivariate multiple regressions correspond to Research Question 2c and provide evidence pertaining to criterion-related validity.

Data Analysis Software

The analyses described above will be accomplished primarily with two types of software. The factor analysis models and model comparisons based on polychoric correlations of attribute ratings will be performed with the EQS 6.3 computer program. The SAS system will perform other correlation analyses, regression analyses, and ANOVAs.

3. Describe methods to maximize response rates and to deal with issues of non-response. The accuracy and reliability of information collected must be shown to be adequate for intended uses. For collections based on sampling, a special justification must be provided for any collection that will not yield "reliable" data that can be generalized to the universe studied.

The Consultation Annual Program Plan (CAPP) included in the annual Cooperative Agreement between OSHA and each state On-Site Consultation program, must contain the projected number of consultation visits that a Consultation program intends to conduct during the fiscal year (FY).

All consultation visits conducted for the Pilot Study will count towards meeting the number of consultation visits projected for the FY, in a state On-Site Consultation program's Cooperative Agreement with OSHA.

In FY 2018, Consultation programs conducted a total of 26,362 consultation visits; in FY 2019, 26,213 consultation visits were conducted; and in FY 2020, 17,663 consultation visits were conducted. Consequently, OSHA is confident about the feasibility of conducting a total of 350 consultation visits (i.e., 100% response rate is expected), projected for the pre-test, Pilot Study, and follow-up study; with Consultation programs applying marketing strategies to inform small- and medium-sized employers in the five (5) industry sectors targeted for these studies, of consultation services available at no cost to them.

Marketing strategies are routinely utilized by state On-Site Consultation programs and fully funded under the Cooperative Agreement with OSHA. These strategies include dissemination of Consultation program newsletters to businesses within the state; presentations at safety and health meetings; and promotion of consultation services on Consultation programs web pages, state social media platforms, and by state agencies such as Workers Compensation Offices. Some states also offer incentives such as workers compensation discounts to employers who use consultation services.

It is important to emphasize that consultants must evaluate every workplace assessed in these studies with regard to all of the 48 DRF33 attributes, which do not explicitly pertain to the multi-employer core element. Also, the four multi-employer attributes must be rated at every workplace with other workers (e.g., temporary, seasonal, and contractual workers) in addition to the host employer's workers. Previous research using the current Form 33 data has illustrated that missing attribute data can severely alter the correlations among scores in the SHP elements with external criteria (Autenrieth et al., 2015).

To avoid DRF33 with missing attribute data, participating consultants will receive comprehensive training and instruction for rating all 52 attributes in the DRF33 and properly completing the worksheet. In addition, the data collection tool will not permit the submission of DRF33 worksheets with incomplete information. If a consultant cannot fully complete a DRF33 for a workplace that DRF33 will not be included in a study, and a replacement workplace will be used instead; by selecting from the pool of employers' requests to Consultation programs for consultation visits. As mentioned earlier, any missing data encountered will likely be due to consultant error when making observations at the workplace or workplace closure due to external events (e.g., the COVID-19 pandemic). Such missing data are unlikely to be informative with respect to the workplace's standing on key study measures, and thus, substitution with a replacement workplace seems reasonable.

Consultants' training will also include a thorough review of each study's procedures and requirements. Similarly, Consultation Program Managers who oversee the daily activities of On-Site Consultation programs will be informed of these procedures and requirements to ensure that they are properly followed.

4. Describe any tests of procedures or methods to be undertaken. Testing is encouraged as an effective means of refining collections of information to minimize burden and improve utility. Tests must be approved if they call for answers to identical questions from 10 or more respondents. A proposed test or set of test may be submitted for approval separately or in combination with the main collection of information.

Pilot Study

A pre-test with fewer than the estimated maximum sample size of 20 was conducted for the Pilot Study following similar procedures for *Design and Sample Characteristics* (see #1), to ensure that all procedures and information technology platforms will function as planned. The pre-test was conducted in only one of the ten regions. Any failings observed in procedures or technology used during the pre-test was corrected before proceeding to the pilot study launch.

A follow-up study with a maximum sample size of 30 will also be conducted for the Pilot Study, following similar procedures for workplace selection and data analyses, when findings result in updates to the DRF33 and ARG that require verification of effectiveness before implementation operationally.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second International Symposium on Information Theory*, (pp. 267–281). Academiai Kiado: Budapest.

Autenrieth, D. A., Brazile, W. J., Sandfort, D. R., Douphrate, D. I., Román-Muñiz, I. N., & Reynolds, S. J. (2016). The associations between occupational health and safety management system programming level and prior injury and illness rates in the U.S. dairy industry. *Safety Science*, *84*, 108-116. doi:10.1016/j.ssci.2015.12.008.

attell, R. B. (1978). The scientific use of factor analysis. NY: Plenum.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Holt, Rinehart, and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334

Finn, J. D. (1974). *A general model for multivariate analysis*. NY: Holt, Rinehart and Winston.

Gibbons, & Hedeker, (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423-436.

Gorsuch, R. L. (1983). Factor analysis (2nd Edition). Hillsdale, NJ: Earlbaum.

Hendrickson, A. E, & White, P. O. (1964). PROMAX: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, *17*, 65-70.

Hu, L., & Bentler, P. M. (1999.) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55.

Joreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika*, *31*,165-178.

Nunnally, J. C, & Bernstein, I. H. (1994). Psychometric theory (3rd Edition). NY: McGraw-Hill.

Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*. 461–464.

Shea, T., De Cieri, H., Donohue, R., Cooper, B., & Sheehan, C. (2016). Leading indicators of occupational health and safety: An employee and workplace level validation study. *Safety Science*, *85*, 293-304.

Tukey, J. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*, 99–114.

5. Provide the name and telephone number of individuals consulted on statistical aspects of the design and the name of the agency unit, contractor(s), grantee(s), or other person(s) who will actually collect and/or analyze the information for the agency.

Patrick A. Showalter is the Director of OSBA. He has the overall authority and responsibility for the Form 33 Revision Project.

Patrick A. Showalter
Director, Office of Small Business Assistance (OSBA)
Directorate of Cooperative and State Programs
Occupational Safety and Health Administration
Washington, DC 20210
202-693-2229; Showalter.patrick@dol.gov.

Patrick Showalter oversees the project manager (Opeyemi Farquah), who is tasked with executing the Form 33 Revision Project. Barney Lawrence functions as Patrick Showalter's alternate.

The project manager works with a team comprised of Paul Schlumper, Brian Warrick, James Robert, Bruce Love, Charles McCormick, Brian W. Sloboda, and Britni Wilcher. Additional subject matter expertise is sought as necessary. Details of Form 33 Revision Project team members:

- Patrick Showalter, Director, Office of Small Business Assistance (OSBA), Directorate of Cooperative and State Programs, OSHA, Showalter.patrick@dol.gov;
- Barney Lawrence, Supervisor, OSBA, Lawrence.Barnett@dol.gov;
- Opeyemi Farquah, M.S., GSP, Safety and Occupational Health Specialist, OSBA, Project Manager, Farquah.opeyemi@dol.gov;
- Bruce Love, MBA, Program Analyst, OSBA, love.bruce@dol.gov;
- Christian Wojnar, Program Analyst, OSBA, Wojnar.christian@dol.gov;
- Brian W. Sloboda, Economist. OSHA, Directorate of Standards and Guidance, Office of Regulatory Analysis-Safety, sloboda.brian.w@dol.gov;
- Paul Schlumper, P.E., CSP, Consultation Program Manager, Georgia Technology Consultation program; Occupational Safety and Health Expert, Paul.Schlumper@innovate.gatech.edu;
- Dr. James Roberts, Ph.D., Associate Professor, Georgia Technology, School of Psychology (Psychometrics/Statistics), james.roberts@psych.gatech.edu; and
- Dr. Brian Warrick, Ph.D., CIH, CSP, Consultation Program Manager, University of South Florida Consultation program; Occupational Safety and Health Expert, warrickb@usf.edu.

For questions about data collection, analyses, and statistical methods, please contact:

- Patrick Showalter, Showalter.patrick@dol.gov;
- Barney Lawrence, Lawrence.Barnett@dol.gov;
- Opeyemi Farquah, M.S., GSP, Farquah.opeyemi@dol.gov;
- Dr. James Roberts, Ph.D., Associate Professor, james.roberts@psych.gatech.edu;
- Paul Schlumper, P.E., CSP, Paul.Schlumper@innovate.gatech.edu;
- Dr. Brian Warrick, Ph.D., CIH, CSP, warrickb@usf.edu.