



# **DATA SUBMISSION PROCEDURES**

**June 2003**

**Report Prepared by CSAP Data Coordinating Center  
Contract No. 277-00-6112**



**U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES**  
Substance Abuse and Mental Health Services Administration  
Center for Substance Abuse Prevention  
[www.samhsa.gov](http://www.samhsa.gov)

# DATA SUBMISSION PROCEDURES

## Table of Contents

1.	Introduction.....	2
1.1	What to send .....	2
1.2	Data submission schedule.....	3
1.3	Data receipt procedure .....	3
2.	Data File Requirements.....	3
2.1	File Format.....	3
2.2	Data media .....	4
2.3	Missing value codes.....	4
2.4	Confidentiality and unique identifier codes.....	5
2.5	Quality assurance.....	5
2.6	Historical values.....	5
3.	Accompanying Documentation .....	5
3.1	Survey instrument.....	5
3.2	Data dictionary.....	6
3.3	File description.....	6
3.4	Record layout.....	6
3.5	Cleaning/editing rules .....	6
	Appendix A: Variable Names and Formats for Items from the GPRA Cross Cutting Tool .....	7
	Appendix B: CSAP Program/PCC Data Submission Cover Page.....	8
	Appendix C: Participant and Site Confidentiality and Publication Agreement Assurances .....	9
	Appendix D: Sample Data Dictionary Entry .....	10
	Appendix E: Sample Record Layout .....	11
	Appendix F: CMI Codebook .....	12

## **1. Introduction**

The purpose of this document is to inform programs/Program Coordinating Centers (PCCs) of the submission requirements for providing data to the DCC in a timely and accurate manner. Included are a timeframe for data submissions, a description of the requirements for each data file, and the necessary documentation to accompany each data submission. A list of names and formats for items from the GPRA cross cutting tool is included in Appendix A of this document. A data submission cover page and data checklist is provided in Appendix B and should accompany each submission. Assurances regarding participant and site confidentiality, as well as publication agreements, are provided in Appendix C. A sample data dictionary entry is contained in Appendix D, and Appendix E presents a sample record layout. A list of names and formats for CMI items is included in Appendix F of this document.

In general, we expect to receive the following categories of data from programs:

- Individual and family participant level data (GPRA, CMI, demographic, and other data)
- Community-level data
- Participant dosage data (method of delivery, duration and frequency of sessions, and other data)
- Program and process level data
- Cost data
- Qualitative data (text, reports) as necessary for interpreting the quantitative data.

Programs will only be required to submit data they are already collecting. They will not be required to collect any additional data as a part of these data submission procedures.

### **1.1 What to send**

The following seven items should be included in every data submission to the DCC:

1. Data file (described in section 2)
2. Data submission cover page and documentation checklist (provided in Appendix B)
3. Survey instrument (described in section 3.1)
4. Data dictionary (described in section 3.2)
5. File description (detailed in section 3.3)
6. Record layout (described in section 3.4)
7. Cleaning/editing rules (described in section 3.5)

## 1.2 Data submission schedule

The DCC requests that data be submitted every 6 months:

- No later than November 30 (last half of the fiscal year).
- No later than May 15 (first 6 months of the fiscal year)

A May data submission permits the DCC to perform centerwide analyses for accountability reports due the following fall. The November data submission period supports centerwide analysis for program planning and the Congressional justification.

Data should be submitted to the DCC on an additive basis; that is, the data submitted in November will not include any of the data previously submitted in May. Programs may submit data following the “natural life cycle” of the program. Where relevant, however, the DCC would like to have data at four collection points: baseline, ongoing (e.g., dosage, process), exit, and 6-month follow-up.

## 1.3 Data receipt procedure

The DCC will confirm receipt of the data with the program GPO within 7 working days of receipt and report any immediate concerns or questions regarding the data based on a cursory review of the files. As data are loaded and prepared for analysis, other questions may arise. The GPO will be notified as necessary.

The DCC will coordinate with the CSAP Project Officers before directly contacting the programs. The Project Officers will provide the DCC with the name of the program GPO. The DCC will send copies to appropriate Project Officers whenever corresponding with the programs/PCCs.

## 2. Data File Requirements

### 2.1 File format

The DCC can accommodate a variety of file formats. Any of the following formats are acceptable:

- **ASCII Text Files.** If the data are submitted in ASCII / text formats, fields should be either fixed or delimited by a comma or a tab.
- **Microsoft Excel Spreadsheets.** The first row in an Excel spreadsheet should contain the column names. The column names should be meaningful for people who are not familiar with the file. There should normally only be one sheet per Excel file. If the Excel file contains more than one sheet, give each sheet a distinctive label so that the DCC can tell them apart. Please delete all sheets that do not contain the specific program data being submitted to the DCC.

- **SAS and SPSS Data Sets.** Variables in SAS and SPSS data sets should have labels. Include any associated format library, if applicable.
- **Microsoft Access Tables.** The column names should be meaningful for people who are not familiar with the file. For instance, a column named “Student Number” is more meaningful than a column named “S\_NO”. There normally should be only one table per database. If the Access database contains more than one table, please give each table a distinctive name so that the DCC can tell them apart. Please delete all tables that do not contain the specific program data being submitted to the DCC.

If data are stored in SQL Server, Sybase, DB2, Informix, Oracle, or similar databases or applications, the data should be exported and saved in one of the above-mentioned file formats before submitting them to the DCC.

## 2.2 Data media

Data may be submitted in any of the following media:

- 3 ½ Diskette
- CD
- Iomega 100 MG zip disk
- Zipped E-mail attachments.

## 2.3 Missing value codes

The DCC will be receiving data from many sources, each with different data collection procedures. In order for the DCC to perform its mission of cross-program analysis, it is helpful to understand the reasons for missing or “blank” field values. There are several reasons for blank or missing values to appear in a data file:

- **Not Collected:** Data are not routinely collected, and the question is not asked. This is usually defined on a site-by-site basis, based on local grantee data map.
- **Not Applicable:** The value is not required, an intentional or logical skip.
- **Missing:** The data are normally collected, but the question was not asked by the interviewer. Additionally, “missing” also could mean that for known or unknown reasons, excluding the above, data that are usually collected are not included in the data file.

Because each of these categories has different implications for data analysis, it is important to distinguish between them. The selection of the code values is left to the existing programs, but the code values and meanings should be submitted with the data dictionaries. Please choose numeric values (-3, -4, -5) for missing numeric variables, or unique characters (#,!, @ ~) for missing character variables.

## **2.4 Confidentiality and unique identifier codes**

In order to protect participant confidentiality, individual identifiers (e.g., name, date of birth) must be stripped from data files prior to their submission to the DCC. Since such individual identifiers cannot be used, unique participants and/or family identifier codes must be inserted to distinguish an individual from the others listed in a data set. These identifiers should be consistent for each individual across all data collection points, across different data types, and across all data submissions, so that data can be grouped by participant and/or family. For example, the unique identifier assigned for data collection instruments for core measures must be the same identifier that was assigned for the GPRA cross cutting tool.

The DCC will honor pre-existing confidentiality agreements held between grantees, PCCs, and CSAP. Assurances regarding participant and site confidentiality, as well as publication agreements, are provided in Appendix C.

## **2.5 Quality assurance**

Programs and the PCCs are responsible for performing quality control procedures on data prior to submission to the DCC. Quality control procedures include accurate file linking, removal of duplicate records, observance of any necessary logic and skip patterns, and completion of appropriate non-response codes. Similar data validity and logic checks will be performed by the DCC upon loading into the DCC data inventory. Error reports will be generated for records that fail to meet minimum quality control criteria. These reports will be sent to the programs for review within a negotiated time frame.

## **2.6 Historical values**

Because participant and some program-level data are collected at several points by the grantees, it is crucial that any values that need to be preserved relative to the collection point are so indicated. If data are submitted in “vertical” fashion, that is, a new record for each data collection point, then the time frame or collection point (initial participation in the prevention program, exit, and/or follow-up) must be included with the participant’s unique identifier in each record. Date of interviews/instrument completion also should be included if available.

## **3. Accompanying Documentation**

Several pieces of documentation are critical to the processing of your data and should accompany each data submission:

### **3.1 Survey instrument**

The survey instrument should be provided in electronic form; if an electronic version of the instrument is unavailable, please include a hardcopy. Please indicate which

cohort(s)/wave(s)/State(s) used the instrument and the date(s) on which the instrument was administered, as well as whether or not the instrument was derived from an already established survey. If the instrument was, in fact, derived from an already established survey, please indicate which one. If the program/survey includes State- or community-level data and all States/communities used the same instrument, indicate any questions asked in one State/community but not in another. If States/communities used different instruments, please provide a copy of each instrument administered.

### **3.2 Data dictionary**

The data dictionary should be provided in electronic or hardcopy form. The dictionary is essentially a list of all the variables on the data file, the instrument item corresponding to each variable, the possible values for each variable, and for discrete variables, the format associated with those values. It is vital that the DCC have a cross-walk between the instrument questions and the variables in the data dictionary. The different types of missing values described above should be clearly identified for each variable. CMI/GPRA questions in the survey should be identified, and the GPRA cross-cutting tool items should use the names provided in Appendix A. CMI items should use the names provided in Appendix F. A GPRA/CMI proxy is a GPRA or CMI item that has been modified in any way; any questions used as proxies for GPRA/CMI items should also be identified. A sample data dictionary entry is provided in Appendix D.

### **3.3 File description**

The file description should include the file name, the year/cohort of the data, and the total number of records in the file. Please indicate whether the data collected is administrative or individual-level data. If the file contains a subset (rather than all) of the data collected, indicate why (i.e., the subset is CMI items only). Please indicate whether the data are a census or a sample, and, if sample, the sampling method that was used. The frequency of data collection also should be indicated.

### **3.4 Record layout**

The record layout should include the variable name, label, length, and whether each variable contains character or numeric data. If the data files submitted are in ASCII format, the record layout also must indicate the starting column for each variable. A sample record layout is provided in Appendix E.

### **3.5 Cleaning/editing rules**

Rules used for cleaning/editing also should be included if the data have been edited or recoded in any way (i.e., items refused by the respondent were filled with “-9”). Indicate to which files (if there are subsets, cohorts, or waves) these rules have been applied.

## **Appendix A: Variable Names and Formats for Items from the GPRA Cross Cutting Tool**

An attached document (in Microsoft Excel) contains the variable names, labels, formats, and response codes that should be used for items from the GPRA cross cutting tool submitted to the DCC. To the extent possible, the names are indicative of the content of the variable. For example, variable HSPLT indicates whether the respondent is Hispanic or Latino; AL30DY indicates alcohol use in the past 30 days.

A similar document for CMI items can be found in Appendix F.

**Appendix B: CSAP Program Data Submission Cover Page**

**CSAP Program/PCC  
Data Submission Cover Page and Documentation Checklist**

Please attach this page with each data submission.

Program \_\_\_\_\_

Program Coordinating Center/Organization \_\_\_\_\_

Address \_\_\_\_\_

Data submitted (inventory all data diskettes/files) \_\_\_\_\_

Reporting Period \_\_\_\_\_

**Point of Contact Information:**

Name \_\_\_\_\_ Position \_\_\_\_\_

Telephone \_\_\_\_\_ E-mail \_\_\_\_\_

Is the point of contact the GPO or the PD? \_\_\_\_\_

Program Project Officer \_\_\_\_\_

Date of Submission \_\_\_\_\_

**Which GPRA cross cutting tool was the reference for your instrument?**

Adult 2002 \_\_\_\_\_

Youth 2002 \_\_\_\_\_

Adult 2005 \_\_\_\_\_

Youth 2005 \_\_\_\_\_

Program Year/Cohort: \_\_\_\_\_

Is the universe a Census? \_\_\_\_\_ or Sample? \_\_\_\_\_

Frequency of Collection: \_\_\_\_\_

Unit of analysis (i.e., individual, program, community): \_\_\_\_\_

Method of administration (i.e., interview, self-report): \_\_\_\_\_

Sampling frame (if applicable): \_\_\_\_\_

**Documentation checklist—did you include:      If no, why not?**

**Survey instruments** \_\_\_\_\_

**Data dictionary** \_\_\_\_\_

**File description** \_\_\_\_\_

**Record layout** \_\_\_\_\_

**Cleaning/editing rules (if applicable)** \_\_\_\_\_

## **Appendix C: Participant and Site Confidentiality and Publication Agreement Assurances**

The purpose of this document is to provide CSAP programs with assurances that the Center for Substance Abuse Prevention Data Coordinating Center (CSAP DCC) will honor and abide existing data sharing publication and transference agreements developed by CSAP Program Steering Committees and approved by CSAP. The CSAP DCC is fully committed to the following assurances:

### **Confidentiality**

The CSAP Data Coordinating Center (CSAP DCC) will maintain the confidentiality of the data provided by each CSAP program. It will abide by CSAP guidelines regarding access, maintenance, and security of data in its possession. Programs will remove identifiers before submitting data to the CSAP DCC.

### **Data Management Assurances**

The CSAP Data Coordinating Center (CSAP DCC) will manage data submitted to the DCC by CSAP programs according to the Federal regulations and SAMHSA guidelines and standard professional practices. CSAP authorized and approved personnel will manage and supervise the data under the direction of the DCC Project Director and CSAP Government Project Officer.

### **Honoring Existing Publication Agreements**

The CSAP Data Coordinating Center (CSAP DCC) will honor and abide by the existing conditions regarding the publication of program data. CSAP DCC will avoid infringement of the agreements. Any proposed DCC publication will be reviewed by each program submitting data for the cross-program analysis.

### **Additional Information**

The Government Project Officer for the CSAP DCC is Beverlie Fallik, Ph.D. Office of Policy and Planning. Questions about these assurances should be directed to Dr. Fallik at: Bfallik@samhsa.gov or 301-443-5827.

## Appendix D: Sample Data Dictionary Entry

Each variable should be listed with its corresponding instrument item, possible values, formats associated with the values, and whether or not the item is a CMI item, an item from the GPRA cross cutting tool, or a CMI/GPRA proxy.

### Example:

Instrument item Q7: RACE

- 1: White
- 2: Black
- 3: Asian
- 4: Native American
- 5: Pacific Islander
- 6: Inapplicable
  
- 9: Not ascertained

This item is a proxy for the race item from the GPRA cross cutting tool.

## Appendix E: Sample Record Layout

The record layout should contain, for each variable, the variable name, label, length, and format (character or numeric). If the data submitted are in ASCII format, the starting column for each variable also must be included. The sample layout below is for an ASCII file; other file formats may eliminate the starting column.

<b>Variable</b>	<b>Label</b>	<b>Starting Column</b>	<b>Length</b>	<b>Format</b>
PERSONID	Identifier	1	8	Character
RACE	Race	9	2	Numeric
GENDER	Gender	11	2	Numeric
AGE	Age	13	3	Numeric

## **Appendix F: CMI Codebook**

An attached document (in Microsoft Excel) contains the variable names, labels, formats, and response codes that should be used for CMI items submitted to the DCC. The CMI domain, sub-domain, construct, and sub-construct numbering system was used to create the codebook to allow for easy reference to the CMI instrument. For example, variable LF01 is listed in the row headed A1\_1\_1, which refers to the first domain (ATOD Use), first construct (Lifetime Use), first scale item (Have you ever smoked cigarettes?).