ASPE Preferences on “Public Use” Data Files

(Revised, June 2000)

The Office of the Assistant Secretary for Planning and Evaluation (ASPE) has provided grants to a number of states and large counties to study outcomes for families leaving or diverted from welfare. Under the grants, awarded in September 1998 and September 1999 (and forthcoming in September 2000), Welfare Outcomes Grantees are required to produce some form of “public use” data files in addition to written reports. A variety of questions about the production and documentation of such files were raised by grantees in March 2000 during a telephone interview/needs assessment conducted by ORC/Macro International, the Technical Assistance contractor for the Welfare Outcomes Grantees. In response to the expressed needs of its grantees, ASPE has undertaken several measures to provide guidance in the production of public use files, including:

- Creation of this short statement of “ASPE preferences” that provides initial responses to questions asked by grantees (draft April 2000, revised June 2000);

- Formation of a Public Use Data Files Work Group, comprised of a small number of grantee and research community representatives, which met in spring/summer 2000 to produce technical guidance on public use data files;

- Development of *Producing and Documenting Data Files: Guidance for ASPE Welfare Outcomes Grantees* (forthcoming), based on the recommendations of the Work Group;

- Distribution of “sample documentation” from one of the grantees (Washington state), prepared following the Work Group’s recommendations in *Producing and Documenting Data Files* (forthcoming);

- Discussion with the National Center for Health Statistics (NCHS) about using its Research Data Center as a mechanism for storing and distributing grantees’ public use data files; and

- Support under an ongoing contract with ORC/Macro International for technical assistance to Welfare Outcomes grantees on issues related to public use files as well as other matters.

In the “Qs & As” below, we provide initial responses to some of the questions raised by grantees in March 2000. More technical assistance, including templates and models, will be provided in the forthcoming guidance on *Producing and Documenting Data Files.* We also encourage grantees to continue to contact ASPE staff and/or the ORC/Macro International staff with additional questions and concerns.

In these efforts, ASPE has been guided by the goal of working collaboratively with grantees to produce data files that will be useful to the research community in its ongoing investigation of welfare outcomes. At the same time, we are firmly committed to producing data files in a manner that protects the identities of individuals who have participated in the Welfare Outcomes studies and does not overburden grantees. While we do not expect all grantees to follow every recommendation of the Work Group, we do expect grantees to review the guidance on *Producing and Documenting Data Files* and to incorporate the Work Group's recommendations when producing and documenting grantee data files, to the extent practicable for the circumstances of their particular projects.

Questions and Answers on Public Use Data Files

A. Data Files to be Produced

Q #1: Should grantees prepare public use data files for both the administrative and survey data used in their analysis?

> A: Yes, ASPE expects grantees to make available both their survey and administrative data for secondary analysis.

Q #2: Should grantees prepare data files with linked survey and administrative data?

> A: Although the April draft of this document originally suggested that grantees prepare a linked file with both survey and administrative data, our revised position, based on the recommendation of the Public Use Data File Work Group, is that grantees produce a set of relational files that can be linked together. As explained further in *Producing and Documenting Data Files* (see sections on "File Structure and Format," and "File Contents"), these files would include a "base" administrative data file, supplemental administrative data file(s); and survey data files(s), which could be all linked together through unique record identifiers.
>
> We recognize that providing linkable files raises some issues. Confidentiality issues are addressed below. In addition, a few grantees used different cohorts for their administrative and survey data. Providing linkable files in such cases may place a burden on the grantee, but we still would suggest that the grantee consider going back to provide a file of administrative data that can be linked to the survey sample. This could be either a small or an extremely burdensome exercise, which may determine its feasibility.

B. General Issues of Confidentiality, Distribution, Researcher Access, Maintenance of Files

Q #1: What guidance can ASPE provide grantees on protecting benefit recipient confidentiality (e.g., federal or other standards, past experience with similar files)?

A: The level of protection within the data file depends heavily on the issue of access. If access is highly restricted in a secure site and signed agreements for the protection of confidentiality are in place, then there is little need to strip data of potential identifiers, such as geography, etc. If the data are on a CD and distributed to anyone who signs a data sharing agreement, the work of stripping and masking personal identities is much more difficult and depends heavily on population and sub population size. Our preference is the former, but we will provide grantees with both options. (See below.)

Further information on confidentiality, including a bibliography of papers related to confidentiality issues in public use and restricted data sets, is being prepared by ORC/Macro International, our Technical Assistance contractor.

Q #2: Is a policy of restricted access consistent with ASPE's definition of "public use" data file?

A: Yes, as suggested by the response above and as stated in the original Request for Applications (RFA), we are using the term "public use" data file in a loose way to refer to files that are made available to researchers for secondary data analysis. Such files may be made available as restricted access data sets in order to ensure the protection of sample members. Steps taken to limit data access, however, should not be so burdensome as to discourage researchers from using the data.

In a survey of grantees conducted in May 2000, we learned that many grantees were in favor of storing the data sets at a secure site and making them available through restricted researcher access. The principal advantage of this approach is that it allows rich data sets with geographic identifiers to be made available to researchers while safeguarding client confidentially. Also, state or county authorization for release of the data may be facilitated if there are assurances that there are safeguards against client identification. In addition, some grantees like the idea of an application process for access to their data. The chief disadvantage of this option is that restricted access places some degree of burden on researchers. Some grantees, therefore, prefer to strip identifying information and produce public use files that do not need to be placed in a secure site. This is also an option, as discussed below.

Q #3: How does ASPE plan to use and distribute the public use data files provided by grantees? What role will ASPE and the grantees play in maintaining the public use data files in the future?

A: ASPE proposes to place the data files at the new Research Data Center operated by the National Center for Health Statistics (NCHS), with two options for access, depending on the confidentiality of the data. If files are designated by the grantee as "restricted access" files, they will be housed in a secure site at the NCHS Data Center. Researcher access to the data would involve making an application for the data and either coming onsite to the Data Center in Hyattsville, Maryland or using a remote access system which is monitored to ensure that there are no breaches in confidentiality or inappropriate disclosure of data. Another option, requested by some grantees, is that the grantee take responsibility for stripping the data files of identifying information and make the resulting “clean” data sets available to researchers, without restricted access. Copies of unrestricted access files could still be stored and distributed through the NCHS Data Center, in order to have all data sets accessible through one centralized location. In addition, grantees may also make the data available themselves.

ASPE proposes to take on primary responsibility for maintaining the files at the NCHS Research Data Center in the future. We propose offering grantees the choice of how involved they would be in reviewing applications for access to their data sets or being notified of such applications. With regard to researchers’ questions about the data files, we propose to establish a centralized email account that would be monitored by NCHS and ASPE staff. Questions that cannot be answered by NCHS or ASPE (or ASPE’s technical assistance contractor) would be forwarded to the grantee. This reinforces the need for grantees to provide well-documented data files, to minimize questions from researchers.

We plan to support the ongoing maintenance of the files at the Data Center through modest researcher access fees. Currently, the NCHS Research Data Center charges $1,000 per week for on-site access, or $500-$1,000 per month for remote access, depending on the size of the files. ASPE plans to subsidize up to 75 percent of these fees, at least initially, to alleviate burden to researchers and encourage access to the Welfare Outcomes data files.

C. Elements of Data Files

Q #1: What survey data should be included in public use data files (e.g, raw data, constructed variables, sample weights)? Should questions not used in analysis be excluded?

A: The Public Use Data File Work Group concurred with ASPE’s basic premise that all survey data collected should be included in grantee survey data files. Just because a question was not used in the analysis does not mean that it should be dropped from the file. Other researchers may want to analyze data that could not be analyzed by the grantee because of time and resource constraints. The general goal should be to provide a file that will provide a rich set of data for future analyses. However, the Work Group did note several important exceptions to this premise. As explained further in the “File Contents” section of *Producing and Documenting Data Files*, these exceptions include such things as “uncoded” open-ended survey questions and obvious personal identifiers.

In addition to the direct survey responses, the Work Group recommends that grantees provide survey weights, as well as certain survey administration items (e.g., interview date, interview mode), as explained further in the “File Contents” section of the guidance.

Q #2: What elements drawn from administrative data sources should be included in public use files (e.g., raw data, constructed variables, and selected elements relating to grantee analysis)?

> A: One of the contributions of the Public Use Data File Workgroup was to develop a set of "common data elements" that are recommended as items to be reported by all grantees. As explained further in the "File Contents" section of *Producing and Documenting Data Files*, these elements include quarterly earnings amounts; flags for monthly receipt of TANF, food stamps, and Medicaid; and key demographic variables. In addition, the Work Group recommends that grantees include any data element that was used in presenting their administrative data reports (referred to by the Work Group as "grantee-specific data elements.") Finally, the Work Group encourages grantees to provide as much administrative data as possible in their data files to support more detailed analysis by researchers. These additional data elements may be provided in supplemental files, linked to the administrative data "base" file through individual record identifiers.

Q #3: How should grantees deal with data elements that are problematic (e.g., survey items with high item non-response, administrative data with missing or inconsistently coded data)?

> A: As stated above, the Work Group concurred with the ASPE preference that grantees provide a full set of data elements. However, if the grantee has some reason to doubt the validity and reliability of a variable, this should be stated in the documentation and code book. Elements that are not reliable for analysis should be documented, with explicit cautions about their use in analysis.

Q #4: What population should be included in public use files drawn from administrative records? Should it be a universe of cases or a sample of cases, and from what time period? Should it include all cases or a subgroup analyzed in the interim and final report?

> A: ASPE prefers a universe of leavers from a specific time that corresponds to the timing of the survey sample cohort. In a few cases, grantees used different cohorts for their administrative and survey data, and so other options may need to be considered. As stated in response to Q#2 in Section A, we still would suggest that the grantee consider going back to their administrative data for the survey sample and linking in the administrative data.
>
> If grantees restricted their analysis to a subset of leavers (e.g., single female parents < 65), yet have a larger group of leavers in their data file, ASPE prefers that the public use file consist of all closed cases, but with a flag to identify the subgroup that was included in the grantee's analyses. This would aid researchers in replicating analyses, but also would allow researchers to expand analyses to additional groups of leavers. In addition, as explained further in the "File Contents" section of *Producing and Documenting Data*

Files, the Work Group recommends that the administrative data files include a flag to identify all survey respondents and non-respondents.

Finally, ASPE prefers data from multiple cohorts, when available.

Q #5: Will there be an effort to standardize data file format and content across grantees?

A: Yes, as stated above, one of the contributions of the Public Use Data File Work Group was to develop a set of “common data elements” that are recommended as items to be reported by all grantees. As explained further in the “File Contents” section of the guidance on *Producing and Documenting Data Files*, these elements include quarterly earnings amounts; flags for monthly receipt of TANF, food stamps, and Medicaid; and certain key demographic variables. In addition, the Work Group proposed a standardized approach to constructing variable names for recommended “common” and “grantee-specific” elements in the administrative data files. For example, TNFR_B01-B12 and TNFR_A00-A12 would be the variable names for receipt of TANF in the 12 months prior to exit, the month of exit and the twelve months after exit. The Work Group also proposed a standardized approach to the treatment of missing values in both the administrative and data files (e.g., using negative numbers between -1 and -9). Finally, the Work Group recommended that all grantees produce data files using an ASCII fixed field file format.

D. File Documentation

Q #1: Should grantees provide written documentation for administrative and survey public use data files?

A: Yes. A principal goal of the Public Use Data File Work Group was to develop standards and a possible model approach to documenting survey and administrative data files.

Two major sections of *Producing and Documenting Data Files* concern “Documenting Administrative Data Files,” and “Documenting Survey Data Files.” As explained in these sections, the Work Group recommends that the documentation of the administrative data files include a narrative (providing general information about the study and files), a data dictionary, a record layout, and certain other items. Likewise, the documentation of the survey data files should include a narrative (with an overview of the survey, description of survey methodology and information about the files), a data dictionary and a record layout as well as certain other items.

To facilitate the development of these documentation materials, ORC/Macro International is preparing some templates, in both Word and WordPerfect, that may reduce the burden of writing the documentation materials. (Use of any of the templates is optional.) The guidance on *Producing and Documenting Data Files* may also include code for constructing a description of record layout from SAS Proc Contents and Put statements.

Finally, one of the grantees, Washington State, has volunteered to “pilot test” the guidance developed by the Work Group by attempting to produce and document its data files using the guidance and templates as they are developed this summer. This sample documentation will be distributed to grantees shortly after issuance of *Producing and Documenting Data Files.*