

NCSES Policy on Matching Non-NCSES Data to Restricted-Use Data Sets

NCSES occasionally receives requests to match NCSES-collected survey data with administrative or other data as a potential research tool. NCSES must review each such request on a case-by-case basis to determine the feasibility, merit and appropriateness of the activity. Any matching with NCSES data will be conducted by NCSES contractors (or, with NCSES approval, by a Federal agency that co-sponsors the survey or that sponsoring agency's approved contractor). The requester will be required by NCSES (or may be required by another sponsoring agency) to pay any costs (to be billed by the appropriate contractor) associated with creating and documenting the matched data set. Such costs may include an up-front fee to the contractor to conduct a needs assessment and develop a full cost proposal.

To preserve NCSES's responsibility to protect the identity of respondents, NCSES reserves the right to require "masking" of data items, or their use only at a secure site if the non-NCSES data comes from a data set, especially one with identifiers, that is available to the requesting researcher. Regardless of the incoming data source, the resulting matched data set, which will not contain any personal identifiers (e.g., names, Social Security numbers), will become an NCSES data set, must be licensed (and might only be available for use at a secure site), will be governed by all regulations pertaining to such data sets, and may be made available to other researchers (NOTE: only through the processing and approval of an NCSES restricted-use data set license for such data).

Proposal for creating a matched data set

NCSES seeks to create the most useful data sets possible and respond to the needs of the user community, but NCSES must balance the merits of each matching request against the need to protect the confidentiality of the NCSES collected data, the potential impact on other NCSES activities, including the use of NCSES resources and/or contractors. Anyone wishing to request a matching of NCSES data with other data must submit a proposal which must contain:

- a description of the data sets to be matched,
- the source and location of the data to be matched,
- verification that the matching is allowed by the producers of the non-NCSES data source,
- the research value of doing the matching,
- how the matched data set will be utilized,
- a description of the characteristics and quality of the non-NCSES data that will be involved in the matching,
- a discussion of the variables and procedures necessary to do the matching, and
- the anticipated results of the matching (positive match rate, false positive rate, proportion of records that can be matched, etc).

NCSES will review the feasibility, appropriateness, and value of undertaking a proposed match, including:

1. Legal issues, potential conflicts of interests for the researcher, and unacceptable disclosure risks that the proposed matched data set may present. Some considerations are:
 - A. Risk of disclosure—a matched data set may substantially increase the risk of disclosure of the respondents' identity,
 - B. To avoid excessive risk, "masking" of the new data may be necessary but excessive masking reduces data utility, and
 - C. Whether the data set can be licensed or can only be used at an NSF, NSF contractor, or survey sponsor site.
2. Assessments of costs and benefits of the proposed match, including but not necessarily limited to:
 - A. Quality of the resulting data set -
 - 1) Description of matching procedures,
 - 2) Identification of issues relevant to the procedures and variables being used, and
 - 3) Anticipated results of the matching (positive match rate, false positive rate, proportion of records that can be matched, etc).

- B. Time constraints in making the match -
 - 1) Availability of the data sets to be matched, and
 - 2) Time schedules of the researcher and of NCSES and its contractor.
- C. Gains from making the match -
 - 1) Merit of the proposed research using the matched data set,
 - 2) Breadth of potential new findings, and
 - 3) Potential for other applications/uses of the resulting data set.
- D. Potential costs of doing the match -
 - 1) Increased demands upon NCSES staff and/or delay/curtailment of other activities,
 - 2) Additional demands upon contractor resources, that could reduce time devoted to NCSES contract activities or the timeliness of such contractor activities, and
 - 3) Potential adverse effects on the affected survey, such as potential for lower response rates.

Documentation of a matched data set

If NCSES approves the proposed match and the requester executes an appropriate contract with the NCSES contractor (or an approved agreement with a sponsoring agency), then the contractor conducting the match will prepare documentation of the match that will contain:

- an assessment of the data quality/characteristics of the non-NCSES data used in the match,
- a description of the matching procedures, and
- an assessment of the quality of matching.

Review of the "new" data

NCSES shall review the above documentation for completeness, shall assure that an appropriate license is in place, and shall review the data set for disclosure issues. NCSES shall provide written approval before any matched data are provided to the researcher under the license.

Using a matched data set

A matched data set may be only used under license, and in some cases only at a secure site. The researcher must clearly understand that the matched data are not the property of the researcher - nor are any abstracts (e. g. subsets, recodes, constructed variables, etc.) of such data. As with all NSF licensed data, a researcher may not attempt to identify any respondent, nor may the researcher match the licensed data with other data. In the case of a matched data set, that includes a prohibition on matching the matched data in the licensed data set back to the original non-NSF data utilized in matching. The researcher must accept and abide by the following additional license conditions for the matched data set beyond the regular license conditions:

1. The researcher shall indicate on all products resulting from the use of the matched data set an NSF quality disclaimer – "NSF does not endorse the non-NSF data utilized in this report, does not assume responsibility for the accuracy of the non-NSF data, and does not necessarily endorse the research methodology used in the report."
2. At the conclusion of the license, the researcher must return all data, including the matched data, to NCSES. No trace of such data may remain on storage media or hard drives of the researcher.

At the conclusion of the project, NCSES will store the matched data, the documentation (especially data quality information), and related data in electronic form as may be necessary for the researcher to comply with any grant requirements for maintenance of records, or replicability of findings.

NSF may allow access to the matched data by other researchers under license, including in some cases at a secure site, provided such researchers agree to all the conditions and demonstrate a need for the merged data set.

This policy regarding matching to NSF restricted-use data sets is provisional, and is subject to change based on experience in implementing the policy.