ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

National Institute of Statistical Sciences Data Confidentiality Technical Panel

Final Report

NCES 2011-608

U.S. DEPARTMENT OF EDUCATION

National Institute of Statistical Sciences Data Confidentiality Technical Panel

Final Report

FEBRUARY 2011

Alan Karr
Education Statistics Services Institute
American Institutes for Research

NCES 2011-608
U.S. DEPARTMENT OF EDUCATION

NISS

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Statistics
Jack Buckley
Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high-priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high-quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public. Unless specifically noted, all information contained herein is in the public domain.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to

NCES, IES, U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

February 2011

The NCES Home Page address is http://nces.ed.gov.
The NCES Publications and Products address is http://nces.ed.gov/pubsearch.

This publication is only available online. To download, view, and print the report as a PDF file, go to the NCES World Wide Web Publications and Products address shown above.

This report was prepared for the National Center for Education Statistics under Contract No. ED-05-CO-0044 with the American Institutes for Research. The content of this report does not necessarily reflect the views or policies of the U.S. Department of Education. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

Suggested Citation

Karr, Alan F. (2011). *National Institute of Statistical Sciences Data Confidentiality Technical Panel: Final Report* (NCES 2011-608). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved [date] from http://nces.ed.gov/pubsearch.

Content Contact

Andrew White
(202) 502-7472
andrew.white@ed.gov

NISS Data Confidentiality Technical Panel: Final Report

Contents

List of Tables

List of Figures

NISS Data Confidentiality Technical Panel: Final Report

This is the final report of the NISS Data Confidentiality Technical panel which was convened by the National Institute of Statistical Sciences (NISS) at the request of the National Center for Education Statistics.

1. Technical Panel Charge and Membership

1.1. The Charge to the Data Confidentiality Technical Panel

The National Center for Education Statistics (NCES) asked the National Institute of Statistical Sciences (NISS) to convene a data confidentiality technical panel (DCTP) to review the NCES current and planned data dissemination strategies for confidential data for the following elements:

- Mandates and directives that NCES make data available.
- Current and prospective technologies for protecting and accessing confidential data, as well as for breaking confidentiality.
- The various user communities for NCES data and these communities' uses of the data.

The principal goals of the DCTP were to review the NCES current and planned data dissemination strategies for confidential data, assessing whether these strategies are appropriate in terms of both disclosure risk and data utility, and then to recommend to NCES any changes that the technical panel deems desirable or necessary.

1.2. Members of the Data Confidentiality Technical Panel

Alan Karr, NISS (chair)
George Duncan, Carnegie Mellon University
Stephen Fienberg, Carnegie Mellon University
Bobby Franklin, Louisiana Department of Education
Gerald Gates, Census Bureau (now, private consultant)
Jerome Reiter, Duke University
Lynne Stokes, Southern Methodist University
Rebecca Wright, New Jersey Institute of Technology (now, Rutgers University)

NISS postdoctoral fellow Anna Oganian provided technical support, including carrying out the experiments described in section 5 of this report.

2. Technical Panel Activities

The DCTP met in Washington, D.C., on December 8, 2006; all members were present (Lynne Stokes joined by teleconference). NCES staff members Paula Knepper and Neil Russell made presentations to the DCTP. Deputy Commissioner Jack Buckley, Chief

Statistician Marilyn Seastrom, and Special Assistant to the Commissioner Andrew White participated in discussions. Subsequent DCTP interactions took place by teleconference and e-mail during the calendar year 2007, and were structured around working groups addressing the following topics:

- *Transformation of the Original Database to the Restricted Database:* Karr, Oganian (see sections 4.4 and 5)
- *Transformation of the Restricted Database to the Public Database:* Duncan, Reiter, Stokes, Wright (see section 4.5)
- *Data Access System Issues:* Fienberg, Franklin, Gates, Karr (see section 4.6)

3. Problem Formulation

The recommendations in section 4 require a concrete formulation of the dissemination problem. Common to all situations considered, and shown in the left-hand panel in figure 1, are:

1. An *original database* $\mathcal{O}$, as collected and edited (for instance, to adjust for nonresponse bias) by an NCES contractor.
2. A *restricted database* $\mathcal{R}$, produced by the data collection contractor from the original database using the NCES DataSwap software (see section 5). Adjustments may be made to maintain consistency with associated universe databases.

Figure 1: Formulation of the dissemination problem

As shown in the right-hand panel in figure 1, there may in addition be

3. A *public database* $\mathcal{M}$ (for "masked") produced from $\mathcal{R}$ by application of one or more methods for statistical disclosure limitation, which is available to the public without licensing (or any other restriction).

Each of $\mathcal{R}$ and $\mathcal{M}$ (if the latter exists) is potentially accessible in two conceptually distinct ways:

1. *Directly*, in the case of $\mathcal{R}$ under license from NCES, and in the case of $\mathcal{M}$ by download from an NCES web site. Any statistical analysis may be performed on either $\mathcal{R}$ or $\mathcal{M}$.
2. *Electronically,* by means of a *data access system* (DAS), to which users submit queries specifying statistical analyses to be performed on $\mathcal{R}$ or $\mathcal{M}$.

As the DCTP understands information received from NCES, NCES is committed to access by license to $\mathcal{R}$ in all cases, and is anxious to provide DAS access to $\mathcal{R}$ and/or $\mathcal{M}$ if confidentiality is not threatened. Consequently, for each NCES data collection, three decisions are necessary:

1. Whether and under what circumstances to allow DAS access to $\mathcal{R}$, as well as the nature of such access.
2. Whether to produce and make available a public database $\mathcal{M}$.
3. If there is a public database $\mathcal{M}$, whether and under what circumstances to allow DAS access to it, as well as the nature of the access.

4. Principal Recommendations

Prior to stating its recommendations, the DCTP strongly commends NCES for the attention and care with which it approaches data confidentiality questions and for its willingness to balance disclosure risk and data utility.

Overall Recommendation: As an overall recommendation, the DCTP strongly recommends that NCES *continue to treat the restricted database $\mathcal{R}$ as "ground truth" in the sense that all NCES analyses and publications are based on it rather than* on $\mathcal{C}$. In particular, this ensures consistency between internal and external analyses.

4.1. Access to Restricted Databases

Recommendation 1: The DCTP recommends that, insofar as practical, *access to restricted databases, whether directly or by DAS, be under license from NCES.*

Elaboration - The resultant structure is shown in figure 2. A DAS accessible only to licensed users has two compelling strengths:

1. A data access system can be of unlimited statistical power, with full scripting capability and multiple user interfaces, including graphical interfaces. In fact, a licensed DAS of unlimited power would obviate the need for physical transfer of data from NCES to licensees, eliminating security and monitoring issues. Of course, a DAS of "unlimited statistical power" might be prohibitively expensive and complex to create and maintain.
2. By recording and analyzing queries processed by the DAS, NCES would have a window into usage to its data that does not exist currently, and which could inform the design and improve the quality of future data collections (see Recommendation 7).

The DCTP acknowledges that a licensed DAS poses issues of authentication and encryption, but believes that current technologies are adequate to deal with them.

This recommendation does not accommodate a public DAS operating on $\mathcal{R}$, which is one access model under consideration—and in one case, in operation—by NCES. The DCTP believes that, given current understanding of the disclosure risks associated with data access systems (Gomatam et al. 2005b; Karr et al. 2006); such a DAS would have to be severely limited in terms of allowable queries and responses to be deemed safe, but might still be feasible.

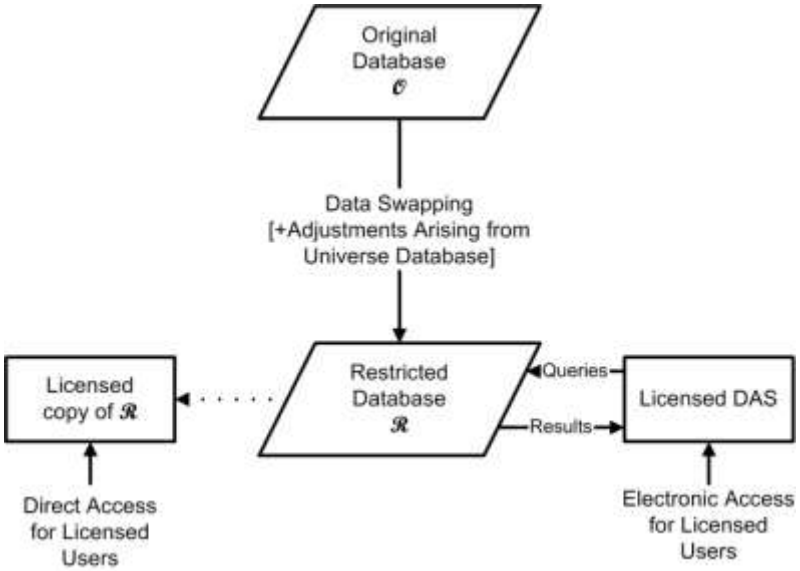
First, the DAS would require query space restrictions in order to address known, related problems, which include the following:

1. *Subsetting of the data.* This is an issue for individual queries; for instance, the mean income of a small number of subjects is more informative about individual incomes than the mean income for a large number of subjects. More subtly (Oganian, Reiter, and Karr 2009), it is also an issue of *query interaction*: by comparing the results of two queries on subsets of the data differing by one

subject, information about that subject is revealed. Preventing the first problem is straightforward;[1] the second is not, since in particular it requires tracking the entire query history for the DAS.

2. *Transformations of variables.* As discussed in Gomatam et al. (2005b), high-leverage transformations of variables entering regressions can reveal individual attribute values. In some ways, such transformations are simply an implicit way of subsetting the data. The query space of a DAS can forbid or severely limit transformations of variables, for instance, by allowing only standard transformations (square roots and logarithms) used to make data "more normally distributed."
3. *Interactions.* While less is known about risks associated with interactions than those arising from subsetting and transformations, there is a problem. Interactions of arbitrarily high order also, in effect, subset the data. The query space of a DAS can limit the order of interactions, although there is no clarity about how much restriction is "enough."

Figure 2: DCTP-recommended structure for programs in which there is *no public database* and only licensed access to the data



Current knowledge is not sufficient to guarantee safety. However, there are two classes of data access systems about which there may be enough known for NCES to proceed:

1. *Table servers* (Dobra et al. 2002). In this case, it is very likely that limiting the dimension (for example, to three or less) of tables provided is adequate to protect confidentiality.[2]

[1] An alternative approach, called *differential privacy*, has been proposed in the computer science literature (Dwork 2007), in which noise is added to query results, and the noise level (variance) is higher the fewer records involved in the query.

[2] Even this is not certain, because methods from computational algebraic statistics can provide information, sometimes very precise, about individual entries in the full table.

2. *Regression servers* (Gomatam et al. 2005b). In this case, of course, restriction of the output is mandatory; for instance, residuals cannot be released. How cautious to be in dealing with the subsetting/transformation/interaction issue is not obvious. However, it is possible that the user community for a public DAS running on $\mathcal{R}$ would be satisfied with extreme caution of the form "no subsetting,[3] no transformations other than square roots and logarithms, and no interactions."

If the query space of a public DAS running on $\mathcal{R}$ is sufficiently restricted, the DCTP urges that NCES consider pre-computing the answers to all—or a large number of—queries, and having the DAS access these rather than $\mathcal{R}$ itself. Doing this reduces security issues arising from public access to a system that interacts directly with $\mathcal{R}$. To illustrate, a table server that provides only tables of dimension 3 or lower from an $\mathcal{R}$ with 100 dimensions would need to pre-compute and store the answers to only approximately 1 million queries.

4.2. Creation of Public Databases

Recommendation 2: The DCTP recommends that NCES *produce public databases whenever possible*.

Elaboration - Given the commitment of NCES to licensing, a public database $\mathcal{M}$ can serve a client base very different from that of $\mathcal{R}$. Users of $\mathcal{M}$ may be relatively unsophisticated or undemanding statistically. For instance, they may seek only tabular or other summaries rather than detailed statistical models; or, they may be students wanting to explore and understand "real data." The implication is that the statistical disclosure limitation (SDL) used to produce $\mathcal{M}$ from $\mathcal{R}$ (see section 4.5) can be rather strong, ensuring that disclosure risk is negligible.

Recommendation 2a: The DCTP recommends that NCES *include weights in all public databases*, perhaps modifying the weights as a precaution.

Elaboration - The DCTP feels that to release $\mathcal{M}$ without weights destroys its utility for any purpose. However, existing SDL theory and methodology have largely ignored the role of weights with respect to disclosure risk.[4] Clearly the values of weights themselves represent disclosure risks. For instance, weights may be informative about geographical detail that has been removed from $\mathcal{M}$. To some extent, the "perhaps … precaution" clause in Recommendation 2a protects NCES against gaps in current knowledge.

4.3. Access to Public Databases

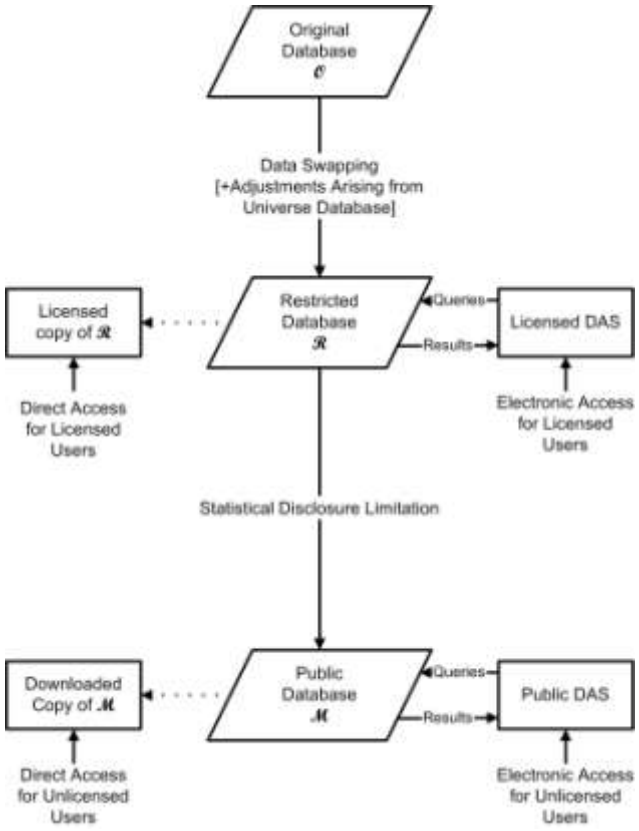
Recommendation 3: The DCTP recommends that NCES *provide DAS access to public databases whenever possible*.

[3] In the case of a database containing both numerical and categorical variables, subsetting only on the categorical variables and only when the associated "cell count" exceeds a threshold.
[4] de Wall and Willenborg (1997) is one exception.

Elaboration - The structure in this case is shown in figure 3. DAS access to $\mathcal{M}$ is not a logical necessity, since anyone who wants to can download $\mathcal{M}$. There are, however, two strong reasons for NCES to provide a DAS:

1. the same "knowing what users want" reason discussed in section 4.1 (see Recommendation 7); and
2. service to clients, since some users of the DAS may not be able to perform the corresponding analyses, no matter how simple, on $\mathcal{M}$.

Figure 3: DCTP-recommended structure for programs *with* public databases



There is, however, a consistency issue discussed further in section 4.5: if Q is a query that a DAS operating on $\mathcal{M}$ will respond to, then the result of Q applied to $\mathcal{M}$ must be identical with the result of Q applied to $\mathcal{R}$.

Neither the query space nor the results provided by public DAS running on a public database $\mathcal{M}$ need be restricted in order to limit disclosure risk, since any user can perform the same analysis on $\mathcal{M}$. Instead, as discussed further in section 4.5, the *inherent nature of $\mathcal{M}$ limits queries and results*.

4.4. Transformation of $\mathcal{O}$ to $\mathcal{R}$

Recommendation 4: The DCTP recommends that NCES *employ the* DataSwap *software to create $\mathcal{R}$ from $\mathcal{O}$*, and maintain its practice of not disclosing associated parameters (attributes, rates, or constraints).

Elaboration - As the DCTP understands it; the rationale for producing $\mathcal{R}$ from $\mathcal{O}$ is to ensure a small level of disclosure protection for all data records with minimal loss of data utility. Data swapping, assuming all records have positive probability of being swapped, is an appropriate means of SDL for accomplishing this, especially for categorical data. In particular, no one-dimensional distributions and most (although data users do not know *which*) multidimensional distributions are not altered.

At the request of NCES, NISS performed extensive computational experiments on DataSwap, which are discussed in section 5. These led to the following subsidiary recommendation:

Recommendation 4a: Because DataSwap shows substantial replicate variability of data utility with other parameters held constant,[5] the DCTP recommends that NCES *run* DataSwap *multiple times on $\mathcal{O}$ to produce multiple candidates for $\mathcal{R}$, and select a candidate with desirable data utility and disclosure risk characteristics*.

4.5. Transformation of $\mathcal{R}$ to $\mathcal{M}$

Recommendation 5: The DCTP recommends that the *public database $\mathcal{M}$ be produced from $\mathcal{R}$ only by means of (1) deletion of (sensitive or other) attributes; (2) category aggregation for categorical variables; and (3) coarsening for numerical variables*.

Elaboration - The central issue associated with transforming $\mathcal{R}$ to $\mathcal{M}$ is consistency between "the same" analyses performed on both. Only licensed users can make such comparisons, but inconsistencies are problematic from almost any perspective. When inconsistencies exist, users only of $\mathcal{M}$ obtain "wrong" answers and worse yet, have no way of being aware of this. Inconsistencies may also be informative about the SDL used to produce $\mathcal{M}$ from $\mathcal{R}$, and therefore subvert the safety of $\mathcal{M}$.

The SDL methods listed in the recommendation enforce consistency, in large part because each restricts the set of analyses that actually can be performed on both $\mathcal{R}$ and $\mathcal{M}$. For instance, a regression involving a variable deleted in $\mathcal{M}$ cannot be performed on both $\mathcal{M}$ and $\mathcal{R}$, so inconsistency is impossible. Similarly, to fit to a categorical database $\mathcal{R}$ the same log-linear model that is fit to $\mathcal{M}$ when $\mathcal{M}$ is produced by category aggregation is impossible without also aggregating within $\mathcal{R}$, and consistency is enforced.

Recommendation 5a: Under some circumstances, subsampling of records may be used in the process of transforming $\mathcal{R}$ to $\mathcal{M}$.

[5] See section 5.3.

Elaboration - The DCTP is not able to offer a definitive recommendation, but does wish to allow the possibility. Two methods of subsampling, which are likely to be used only in conjunction with the SDL methods listed in Recommendation 5, were identified:

- Deleting particularly risky records, assuming that a record-level measure of risk is available. This method has some possibility of being informative about the SDL that transforms $\mathcal{R}$ to $\mathcal{M}$, but introduces only deviations rather than inconsistencies, since if records in $\mathcal{R}$ and $\mathcal{M}$ are not in one-to-one correspondence, then carrying out the same analysis on both is not possible.[6] However, constructing appropriate record weights for $\mathcal{M}$ may be difficult, especially if high weight records in $\mathcal{R}$ are risky.
- Randomly sampling, without replacement, records from $\mathcal{R}$ to include in $\mathcal{M}$,[7] with sampling probabilities potentially reflecting record weights in $\mathcal{R}$. This method also seems unlikely to be informative about the transformation of $\mathcal{R}$ to $\mathcal{M}$, and introduces deviations rather than inconsistencies. Constructing appropriate record weights for $\mathcal{M}$ may be feasible.

The DCTP anticipates that (under circumstances that are virtually impossible to characterize *a priori*) NCES would trade off subsampling against the other three methods in Recommendation 5. For instance, variables that category aggregation might render effectively useless might be "rescued" by subsampling. How NCES would actually perform such tradeoffs is not clear, however.

Current SDL technology does not allow straightforward characterization of additional risk if $\mathcal{M}$ has been constructed from $\mathcal{R}$ by subsampling and weights have been adjusted.

Recommendation 5b: The DCTP recommends that NCES *provide users explicit information about SDL strategies used to transform $\mathcal{R}$ to $\mathcal{M}$.*

Elaboration - Unlike the situation addressed in Recommendation 4, where releasing the identities of the swap attributes or the value of the swap rate poses a threat to confidentiality, releasing such information as "The employer type attribute was deleted" or "Age was aggregated from age in years in $\mathcal{R}$ to 5-year intervals in $\mathcal{M}$" poses no threat at all. It seems likely that the information about $\mathcal{R}$ is already public, so the thrust of this recommendation simply is that NCES serve users by providing the information to them rather than making them locate it themselves.

The issues are subtler in the case of subsampling, especially if it is performed in order to remove risky records. If random subsampling is performed, releasing the sampling rate seems safe.[8]

---

[6] There is a subtle issue here: when results from $\mathcal{M}$ deviate from results from $\mathcal{R}$, even if there is no inconsistency, there is potential for confusion. The results from $\mathcal{R}$ are "true" (as population-level estimates, e.g.,) but those from $\mathcal{M}$ are not. NCES would face the issue of ensuring that users of $\mathcal{M}$ understand this.
[7] As is done, for example, by the Census Bureau to produce Public-Use Microdata Samples (PUMS).
[8] This is done for the Census PUMS.

Recommendation 5c: The DCTP recommends that, in order to inform its choice of SDL strategies to transform $\mathcal{R}$ to $\mathcal{M}$, NCES *explore, and if feasible implement, processes that lead to more precise knowledge of the user community and uses for* $\mathcal{M}$.

Elaboration - Underlying much of this section is a sense by the DCTP that NCES may not have an entirely clear or complete understanding of the user community, and hence the main uses of $\mathcal{M}$, both of which influence significantly the choice of SDL strategies. For instance, if $\mathcal{M}$ were used primarily for "looking up fast facts" and producing cross-tabulations at high levels of aggregation, then category aggregation may be the preferred strategy. By contrast, if the primary use of $\mathcal{M}$ were for exploratory data analysis (in courses, for example), random sampling might be the preferred strategy, especially if it obviates the need to remove variables of scientific interest.

The process of developing better knowledge can, for instance, include polls of users or using information collected by a public DAS (see Recommendation 7).

4.6. DAS Considerations

Recommendation 6: The DCTP recommends that NCES *tailor the user interfaces (UIs) of data access systems to user communities*, to the extent that NCES understands these communities.

Elaboration - By UIs, the DCTP means

- methods for providing inputs to the DAS; and
- mechanisms by which the DAS presents output.

In the setting of figure 2 or figure 3, because restrictions on analyses are imposed by the nature of $\mathcal{M}$, the computational engines underlying the licensed DAS and the public DAS can be identical. Performance or other (see below) reasons might dictate otherwise, however, especially since the user community for a public DAS is almost certainly larger than that for a licensed DAS.

The UI for a licensed DAS can presume some level of technical sophistication of users, and may be more script- rather then point-and-click- oriented. Indeed, if the recommendations in section 4.5 were implemented, the licensed DAS could, at the extreme, not offer any functionality offered by the public DAS, since licensed users can also use the latter. The DCTP feels that it is essential that a licensed DAS provide scripting capability that supports complex, user-specified analyses and linking of multiple analyses, as well as modeling not built into the DAS. A licensed DAS should also provide interactive graphical and text output, as well as export capability for both.

On the other hand, the UI for a public DAS may be highly menu-driven, so that, for example, users can form a cross-tabulation merely by clicking on the variables that define it. The set of allowable analyses in a public DAS may be restricted[9] in order to make the

[9] As discussed previously, this restriction is not necessary in order to protect confidentiality, since users of the public DAS can always download $\mathcal{M}$ and perform any analysis.

UI more user-friendly (more comprehensible, easier to navigate, etc.). Graphical and text output may be designed to maximize understandability rather than maximize detail. Scripting seems not merely unnecessary, but actually gratuitous.

The issues just discussed presume knowledge of uses for $\mathcal{R}$ and $\mathcal{M}$ that may not currently be readily available. The final recommendation addresses this.

Recommendation 7: The DCTP recommends that NCES *configure all data access systems to collect as much information about database uses as possible*, and that NCES use such information to inform modification of existing data products and design of future ones.

Elaboration - The DCTP is aware that collecting such information raises issues, but believes that these can be addressed. The need is for summary information about what kinds of analyses are performed, involving which variables, and with what frequency. There is no need for individually identifiable information, or to look at the pattern of analyses performed by any single user.[10] A simple logging capability would suffice, and would be no more intrusive than web server logs that count page hits.

[10] The DCTP notes that there are reasons such a capability may be desirable in some settings. For instance, in a DAS running on restricted data that are not also available under license, tracking queries from individual users would assist in identifying linked queries aimed at breaking confidentiality rather than legitimate use of the data.

5. The DataSwap Software

NCES provided NISS a copy of its DataSwap software, which is used to produce the restricted database $\mathcal{R}$ from the original database $\mathcal{O}$. The principal purpose was to enable NISS to study this process. Whether the process offers "sufficient protection" in terms of disclosure risk can ultimately be decided only by NCES. Based on the NISS study of DataSwap, there is no evidence that it performs other than as stated, and no evidence that it cannot meet NCES requirements. The experiments performed by NISS focused on exercising DataSwap in order to develop a more detailed understanding of its behavior They are reported in sections 5.2 and 5.3.

5.1. General Comments

DataSwap is a powerful software tool for swapping one or multiple attributes in categorical databases in which records carry weights necessary for analyses. NISS was provided version 2.0.2 of DataSwap, which is implemented as a macro for SAS version 8. This version of DataSwap would not operate under version 9 of SAS—the most recent version, which caused non-trivial delays in performing the experiments. NCES needs to be aware that this may become a more serious issue in the future, although possibly NISS did not have the most recent version of DataSwap.

The strengths of DataSwap include

- Accommodating weights in both selection of swapped records and swap partners
- Algorithms that favor swapping records with those in "adjacent" cells in the contingency table, in order to minimize distortion of the data
- Ability to handle hard constraints representing values that cannot be swapped
- A balanced mode that equalizes swapping across attributes when multiple attributes are swapped

Weaknesses of DataSwap, some of which are shared by other software packages for data swapping,[11] are the following:

- DataSwap does not appear to automatically test whether each individual swap is a true swap, in the sense that it changes the data.
- DataSwap does not appear to detect more complicated compensating multiple swaps that leave the data unchanged.
- DataSwap appears to have no quantified measures of disclosure risk that can be used to compare the pre-swap (original) database $\mathcal{O}$ and the post-swap database $\mathcal{R}$.
- DataSwap provides some quantification of differences between $\mathcal{O}$ and $\mathcal{R}$, but there seems to be no capability to compare the utility—in the sense, for instance, of Gomatam et al. (2005a) or Karr et al. (2006)—of $\mathcal{O}$ to that of $\mathcal{R}$. Sections 5.2 and 5.3 focus on this question.

[11] Such as the publicly available NISS Data Swapping Toolkit (NISS, 2003).

As a result of items 1 and 2, the amount of swapping that actually takes place may be less than the swap rate alone suggests. This may be a concern for NCES if the swap rate is believed to be the primary determinant that disclosure risk is acceptably low.

NISS also reviewed, and finds excellent, the user documentation (Westat 2005) for DataSwap. Once SAS version problems were resolved, NISS was able to run DataSwap without any difficulty.

Were NCES to produce another version of DataSwap, possible improvements include breaking the (seeming) tie to a specific version of SAS, checking that all swaps are "true," and incorporation of utility measures, such as Hellinger distance (section 5.2), with a basis in the SDL literature.

5.2. Data Utility

In this section and the next, data utility is operationalized as the distortion between the original database $\mathcal{O}$ and the restricted database $\mathcal{R}$, as measured by the *Hellinger distance* between the associated contingency tables. This distance has been employed in Gomatam, Karr, and Sanil (2005a); Karr et al. (2006); and elsewhere as a broad but blunt measure of data utility.[12] Given tables T_1 and T_2,[13] the Hellinger distance between them is

$$HD(T_1, T_2) = \frac{\sum_c \left(\sqrt{T_1(c)} - \sqrt{T_2(c)}\right)^2}{\sum_c T_1(c) \sum_c T_2(c)}$$

In this equation c indexes the cells of the tables. This measure emphasizes differences in small-count cells. The results of the experiments are values of $HD(\mathcal{O},\mathcal{R})$ for possibly multiple versions of $\mathcal{R}$.

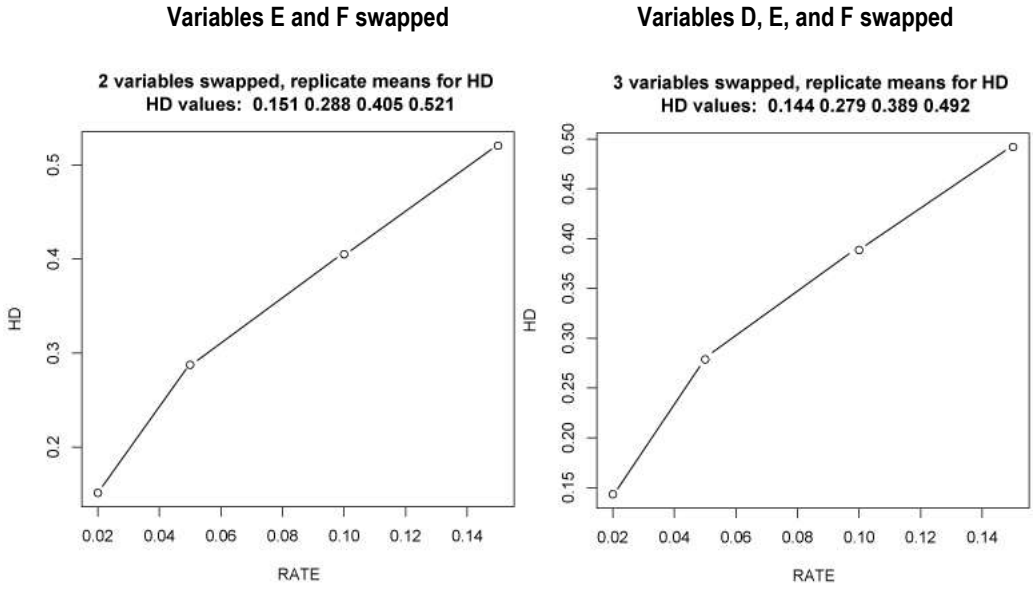
The experiments were carried out on the "Czech auto worker" database (Dobra et al. 2002; Edwards 1985), which consists of 1,841 records containing 6 binary, work- and health-related variables for workers in an automobile factory in Czechoslovakia. The database is listed in tabular form in appendix A. Three of the variables can be construed as public: A = "smokes," B = "performs strenuous mental work," and C = "performs strenuous physical work." The other three variables are private, and need protection: D = "high systolic blood pressure," E = "high ratio of β to α lipoproteins," and F = "family history of coronary heart disease."

Figure 4 shows the behavior of mean data utility (mean Hellinger distance, so that low values represent high utility) averaged over 100 replicates—that is, 100 runs of DataSwap with the same parameters but different random number seeds—as a function of the swap rate, which takes values 0.02 (2 percent of the data are swapped), 0.05, 0.10, and 0.15. In the left-hand panel, variables E and F are swapped; in the right-hand panel,

[12] The propensity-score-based measure proposed in Woo et al. (2007) was not investigated, but may also be relevant. This measure gauges the extent to which $\mathcal{O}$ and $\mathcal{R}$ are indistinguishable.
[13] These can be interpreted as arising from $\mathcal{O}$ and $\mathcal{R}$, respectively.

all three private variables (D, E, and F) are swapped. The results are as expected and consistent with Gomatam, Karr, and Sanil (2005a): the amount of distortion is an increasing, concave function of the swap rate, but since the rates in figure 4 are small, the deviation from linearity, while evident visually, is not dramatic. For each rate, there is little difference in distortion between the 2-variable and the 3-variable case. This suggests that NCES may be able to swap fewer variables than might have been thought necessary. Alternatively, to the extent that swapping more variables reduces disclosure risk, figure 4 suggests that there may be little data utility penalty for doing so.

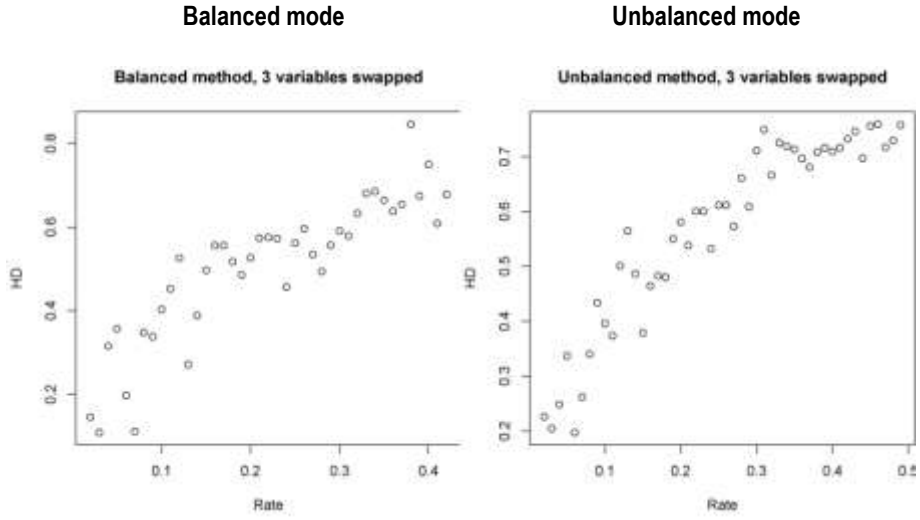
Figure 4: Mean Hellinger distances, averaged over replicates differing only in the random number seed, for various swap rates



NOTE: Lines are added for clarity, and do not represent experimental results.

Figure 5 is a comparison of the utility measure for balanced (left-hand panel) and unbalanced (right-hand panel) modes of DataSwap for swapping rates from 0.01 to 0.5. (For balanced swapping, it was not possible to complete the swapping successfully for rates exceeding 0.42.) In figure 5, each point corresponds to only one replicate, rather than to a mean over multiple replicates, as in figure 4, and variables D, E, and F were swapped. While there are no striking differences, it does appear that at (possibly unrealistically) high swap rates, unbalanced swapping engenders greater distortion.

Figure 5: Comparison of swapping modes



NOTE: Each point corresponds to one replicate. Variables D, E, and F were swapped.

5.3. Replicate Variability

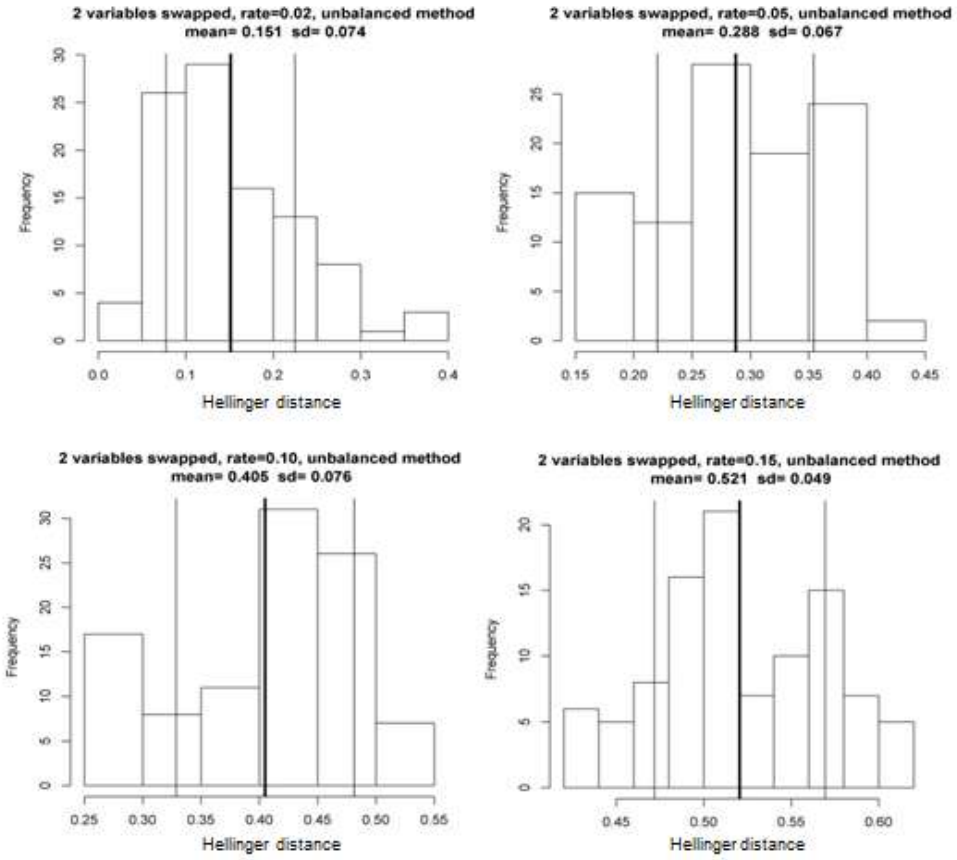
As noted in section 5.2, the points in figure 5 represent one replicate per swap rate. If the points instead depicted averages over multiple replicates, as in figure 4, the results would have been much smoother. But then, figure 5 raises the issue of replicate variability in DataSwap. Figures 6 and 7 address this issue for 2-variable and 3-variable swaps. Each histogram summarizes the results from 100 replicates that have identical parameters other than the random number seed.

Table 1: Means, standards and coefficients of variation derived from Hellinger distances whose histograms are shown in figures 6 (2 variables) and 7 (3 variables)

2-variable swaps (figure 6)			
	Mean	Standard Deviation	Coefficient of Variation
Swap rate = 0.02	0.151	0.074	0.490
Swap rate = 0.05	0.288	0.067	0.233
Swap rate = 0.10	0.405	0.076	0.188
Swap rate = 0.15	0.521	0.049	0.094
3-variable swaps (figure 7)			
Swap rate = 0.02	0.144	0.073	0.501
Swap rate = 0.05	0.279	0.055	0.197
Swap rate = 0.10	0.389	0.051	0.131
Swap rate = 0.15	0.492	0.054	.110

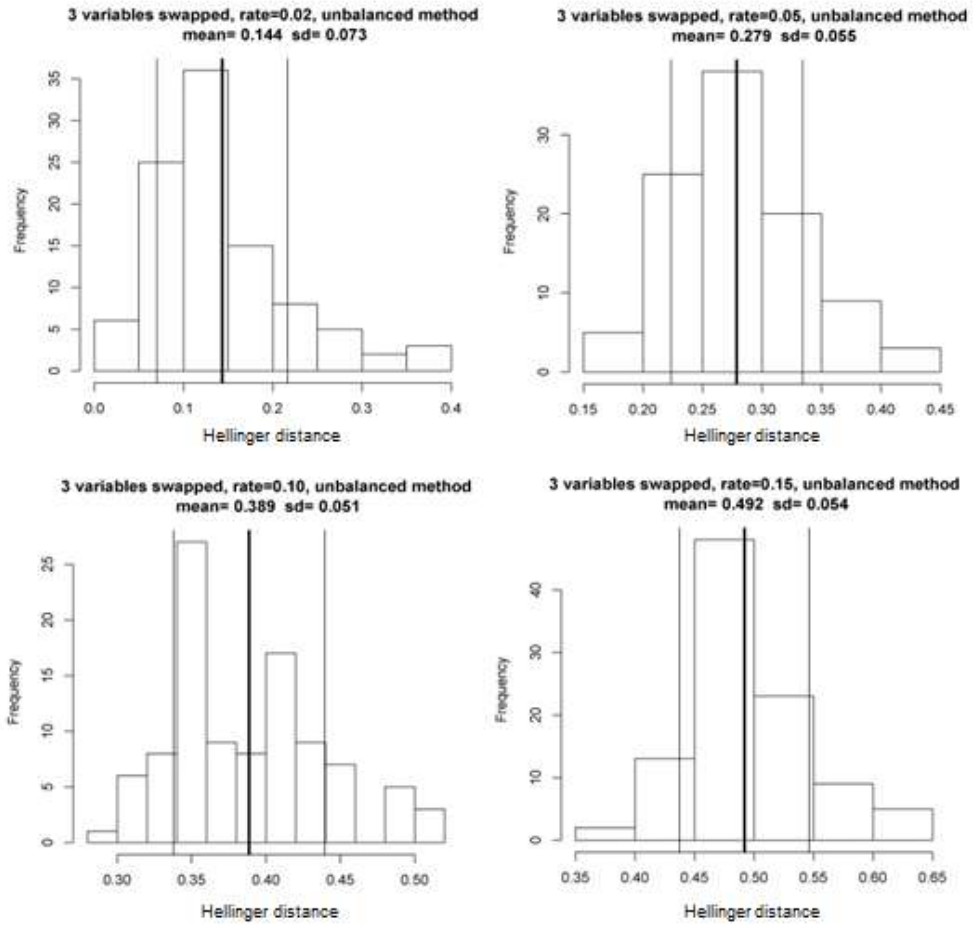
In contrast to the NISS Data Swapping Toolkit, which was used to produce the results in Gomatam, Karr, and Sanil (2005a), where replicate variability was examined but not seen to be an issue, DataSwap exhibits significant replicate variability, which is relatively higher at lower swap rates. Table 1 summarizes the means, standard deviations, and coefficients of variation from figures 6 and 7.

Figure 6: Histograms showing replicate variation of Hellinger distance for *2-variable swap* and multiple swap rates



NOTE: Each histogram corresponds to 100 replicates. *Upper left:* swap rate = 0.02. *Upper right:* swap rate = 0.05. *Lower left:* swap rate = 0.10. *Lower right:* swap rate = 0.15.

Figure 7: Histograms showing replicate variation of Hellinger distance for *3-variable swap* and multiple swap rates



NOTE: Each histogram corresponds to 100 replicates. *Upper left:* swap rate = 0.02. *Upper right:* swap rate = 0.05. *Lower left:* swap rate = 0.10. *Lower right:* swap rate = 0.15.

This replicate variability poses both an important question and an opportunity for NCES: is there corresponding replicate variability in disclosure risk? Neither the disclosure risk measure used in Gomatam, Karr, and Sanil (2005a), namely the percentage of unswapped records in small count cells in $\mathcal{R}$, nor the measures employed in Gomatam et al. (2005b) and Karr et al. (2006), which are specific to numerical rather than categorical data, makes sense in the context of these experiments. The former might be applicable to NCES tables that are larger and hence sparser, but determining this would require further investigation. The implications, however, are clear and divergent:

- If there is *not* corresponding variability in reasonable measures of disclosure risk,[14] then the opportunity alluded to above arises. From multiple replicates generated by DataSwap, all of which are candidates for $\mathcal{R}$, NCES can pick the

[14] This would be true, for instance, if NCES simply declared that the swap rate *is* the disclosure risk measure.

one with highest utility, i.e., lowest Hellinger distance from $\mathcal{O}$. The histograms in figures 6 and 7 show that, especially for lower swap rates, this can yield a dramatic increase in utility at no cost in risk.

- If there *is* corresponding variability in reasonable measures of disclosure risk, then the swap parameters alone do not determine disclosure risk. Therefore, NCES cannot be assured that any given replicate will possess acceptably low risk, and must test candidates for $\mathcal{R}$ individually to determine if the risk is acceptable. The opportunity remains, however, albeit in altered form: from multiple replicates possessing acceptable risk, NCES can then select the one that has the highest utility (lowest Hellinger distance from $\mathcal{O}$).

Empirical evidence in Gomatam, Karr, and Sanil (2005a) and elsewhere indicates that disclosure risk and data utility are strongly related, to the extent that one can often serve as a surrogate for the other. In the context of the decision problem of selecting $\mathcal{R}$, this could provide a short-term solution: select as $\mathcal{R}$ the replicate whose Hellinger distance from $\mathcal{O}$ is minimal, but above a prespecified threshold.

6. Other Database Issues

Two other issues, both that are likely to pertain to all federal statistical agencies, were raised in discussions among DCTP members, who did not come to grips with either issue to the point of producing recommendations for NCES. The issues, which have a common thread of how to do SDL across multiple data releases, are

1. *Longitudinal databases*. Concerns arise from adding both additional attributes to existing cases and freshening, as well as the extent to which SDL on early data releases limits that on later releases.[15]
2. *Multiple databases*. An excessively "stove-piped" approach by NCES may fail to address confidentiality issues arising from linking multiple databases. This issue may be most acute between surveys and universe data collections: can, for example, schools in a survey be re-identified by linkage to the Common Core of Data (CDD), Private School Survey (PSS), or Integrated Postsecondary Education Data System (IPEDS)?

The issues are not associated solely with risk: "coordinated" SDL may also have significant data utility implications.

[15] For instance, running DataSwap independently on a first wave release and then on a second-wave release would enable detection of swapped records by comparing the two databases. This can be avoided if, as may be likely, the swap variables are "quasi-identifiers" already present in the first wave, but then there is no protection from incremental risks posed by additional data.

References

1. A. G. de Wall and L. C. R. J. Willenborg (1997). Statistical disclosure control and sampling weights. *J. Official Statist.* 13(4) 417-434.

2. A. Dobra, S. E. Fienberg, A. F. Karr, and A. P. Sanil (2002). Software systems for tabular data releases. *Int. J. Uncertainty, Fuzziness and Knowledge Based Systems* 10(5) 529-544.

3. C. Dwork (2007). Differential privacy. Available on-line at research.microsoft.com/research/sv/DatabasePrivacy/dwork.pdf.

4. D. E. Edwards and T. Havraneek (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika* 72 339-351.

5. S. Gomatam, A. F. Karr, and A. P. Sanil (2005a). Data swapping as a decision problem. *J. Official Statist.* 21(4) 635-656.

6. S. Gomatam, A. F. Karr, J. P. Reiter, and A. P. Sanil (2005b). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statist. Sci.* 20(2) 163-177.

7. A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* 60(3) 224-232.

8. National Institute of Statistical Sciences (2003). Data Swapping Toolkit. Software and documentation available on-line at www.niss.org/software/dstk.html.

9. A. Oganian, J. P. Reiter, and A. F. Karr (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics and Data Analysis* 53(4) 1475-1482.

10. WESTAT, Inc. (2005). DataSwap User's Guide, Version 2.0.

11. M. -J. Woo, J. P. Reiter, A. Oganian, and A. F. Karr (2007). Global measures of data utility for microdata masked for disclosure limitation. To appear in *J. Privacy and Confidentiality*.

Appendix A: The Czech Auto Worker Database

				B	no		yes	
F	E	D	C	A	no	yes	no	yes
Neg	< 3	< 140	no		44	40	112	67
			yes		129	145	12	23
		≥ 140	no		35	12	80	33
			yes		109	67	7	9
	≥ 3	< 140	no		23	32	70	66
			yes		50	80	7	13
		≥ 140	no		24	25	73	57
			yes		51	63	7	16
pos	< 3	< 140	no		5	7	21	9
			yes		9	17	1	4
		≥ 140	no		4	3	11	8
			yes		14	17	5	2
	≥ 3	< 140	no		7	3	14	14
			yes		9	16	2	3
		≥ 140	no		4	0	13	11
			yes		5	14	4	4

The three "public" variables are A = "smokes," B = "performs strenuous mental work," C = "performs strenuous physical work," and the three "private variables" are D = "high systolic blood pressure," E = "high ratio of β to α lipoproteins" and F = "family history of coronary heart disease." The table is reproduced from Dobra et al. (2002).