The Researcher's Practical Guide to Clearance of Output
(particularly for LRD research)

by Ethan Lewis
updated by Philip Doctor
December 2003

The purpose of this document is to describe, from the point of view of you the researcher, what is necessary in order to bring research output out of the RDC. It should be noted from the beginning that the technical details of these requirements are described in the RDC "Handbook for Researchers" (hereafter referred to as "the RDC handbook.") which you should also read. This guide supplements the handbook by providing a practical description of the steps involved in getting output cleared.

I work with LRD data, so this document is particularly relevant to how the clearance process works for LRD research results. There are three requirements for clearing research results that make use of LRD data:

1. Fill out the disclosure form. It provides documentation of how and where you did your calculations, and how this relates to your project as originally proposed.
2. Provide evidence that the research results do not disclose information on an inappropriately small number of firms. The LRD rules for this are very specific, and provided below.
3. Provide a benefit to the Census Bureau.

1. Filling Out Disclosure Forms

What I will call the non-disclosure form is referred to in the handbook as the template for the "project clearance memo." You can get this form from the RDC administrator. It seeks documentation of how and where you did your calculations to produce the proposed research output, and how the research output being requested is related to the project as originally proposed.

In this section, I will attempt to explain how each question on this form should be answered. I am aware, however, that each researcher's situation may be unique so my descriptions may be sometimes inadequate. CES wants you to fill out the form in a clear enough manner so that given your descriptions, they could repeat what you did and replicate your results. Another important way to think of it is that you should put on this form whatever will make it easier for the RDC administrator (or, say, someone at CES reading your form) to determine if your output is safe. Give clear and organized descriptions of your output and its disclosure safeguards.

(i) At the top of the form, put the date of submission.

1. *General Information.*

This section describes the disclosure request in general, and how it relates to the project.

a. Name of the subdirectory… Just give the filename and path of the electronic copy of the disclosure form you are submitting to the RDC administrator (eg, /rdcprojects/ch/ch00985/disclosure/October2003/clear01.doc).
b. Outputs proposed for removal: Give the name of the electronic files you want to take out. Then describe what they contain. Here the main point is to mention how confidential data was used to create the proposed output (e.g., specify that "Title 13/26 protected capital stock data were regressed on non-confidential geographic variables…"). I would like to mention here that all the statistics you wish to take out must not risk disclosure of confidential information.
c. Part of the approved project. Tell how these results that you want to take out are related to your proposal. If your proposal posed some question, for example, tell how these results answer that question.
d. Use of output: check off all that apply.

2. *Descriptions of Research Samples*.

This section will ask you several questions about your data samples that underlie your output. You will need to start by briefly describing each data sample that you've used. After you've done this you must also describe any relationships between the samples along with any relationships between the samples and other data samples which might be similar.

3. *Output Files*.

This section draws a distinction between "research" and "disclosure" output files. The "disclosure" output files are the ones where you provide documentation that your research output meets non-disclosure requirements and the tables that provide disclosure statistics. "Research" output files are the output files you propose to remove from the RDC (for example the SAS ".lst" files).

In this section you will have to detail information on your research output files. Each research output file should have a supporting disclosure analysis output file containing statistics that support your request for disclosure. For each of these files you should enter:

- File Number. This is nothing more than an ordinal numbering of all the files (1,2,3…) you are listing.
- File Name. Name of the file *and* its location (eg "/rdcprojects/ch/ch00985/disclosure/October2003/Output.log"). Note that it may be convenient instead to just to give the location of *all* of the files listed above in the table if they are all stored in the same location (eg "all files are in "/rdcprojects/ch/ch00985/disclosure/October2003/" then when listing the file names simply enter "Output.log").
- File Description. Describe the sample you used (eg, "large northeastern manufacturing firms") and the disclosure output in this research file. (eg, "regressions of capital stock on employment.")

- Research Output Program. The program that produced your research output. Please list its name and location (same as for the Research Output File's name).
- Research Sample Number. This is the sample number that underlies the research output.
- Disclosure Analysis File Name. This is the name and path of the file containing supporting statistics that allow your research output to be disclosed.
- Disclosure Analysis Program. This is the name and path of the file used to generate your Disclosure Analysis statistics.
- Results. Leave this area blank, CES fills in whether or not this file was cleared.
- Comments. Describe how confidential data sets were used to generate this file, and a general description of how this meets disclosure requirements, (eg, by involving a large number of firms) and any other information that would aid the administrator in assessing the disclosure risk.

4. *Variable Definitions*.

This section asks you for information regarding every variable in your research output. For each variable please list:

- Variable Name. This is simply asking you to indicate the name of the variable you are about to discuss.
- Definition. Write a brief description of the variable (e.g., variable "c0192" contains the cost of cement). If you generated this variable yourself, or if this is a modified variable rather than one that was present in the original data set(s), please indicate the equations/methods used to generate this variable.
- Source. List which data set this variable came from, or if you generated it yourself, please indicate this.
- Comments. Please leave any comments that you feel would help determine if the variable meets disclosure requirements or would help someone follow the logic of your program(s).

5. *Other Information or Comments*.

Anything else the administrator should know. Once again, think of any information that you think might make it easier for the RDC administrator to easily determine if your output is safe for clearance.

2. Demonstrating Non-Disclosure

For any tables or regression output you wish to take out of the RDC, you must produce a series of corresponding tables containing non-disclosure statistics. ***Demonstrating non-disclosure to the RDC administrator is your responsibility as a researcher***. Providing clear documentation and easy explanations for what you do will in general allow your output to be cleared much more quickly. However if your documentation is insufficient and the output appears to disclose sensitive information, your clearance request will be denied, regardless of whether or not you are actually disclosing sensitive information in your output. This section describes some of the specifics of how to do this, but you should also refer to the RDC handbook.

The disclosure requirements are aimed at preventing variables that might disclose sensitive firm level information from being revealed to the public or competing firms.

Demonstrating that your results meet non-disclosure requirements in general terms, has two pieces:

1. Showing that a sufficiently large number of firms were involved in the calculations; and
2. Showing that shipments (TVS) are not too highly concentrated among a small number of firms involved in the calculations.

Before I reveal the specifics of these requirements, I want to emphasize two things. First, it is the number of *firms* not establishments that is the relevant unit for LRD disclosure rules. This is important because the LRD is an establishment level survey, and most researchers work with establishments as the unit of observation. For disclosure analysis, however, these establishments must be collected into firms (how this is done is described in the SAS appendix below). Second, you must show that *shipments* are not too concentrated, regardless which LRD variables you personally are working with.

Now let's consider the specific rules. The handbook distinguishes disclosure rules for "tabular" output (eg, summary statistics) and regression output. However the rules for regression output can actually be thought of as a special case of the rules for tabular output. Hence I will begin with what you need to do for tabular output and then consider regression output.

Tabular Output
It is easiest to demonstrate non-disclosure by creating a table containing all of the descriptive statistics that you wish to clear. Then for each of these statistics show that:

1. The number of firms used to create that cell was not less than XXX.
2. The top YYY firms in terms of shipments (TVS) produce no more than ZZZ percent of the shipments of all the firms used to create that cell. While you only need to computer one ratio, many RDC administrators prefer for you to this for 3 different YYYs. These ratios of shipments from the top firms to the total shipments of your sample are referred to as "Concentration Ratios".

(**Please note that these specific rules are confidential…your administrator will verbally tell you X, Y, & Z**.)

Given what needs to be shown, it should make sense to organize your disclosure output as a table. One column can hold your descriptive statistics that you wish cleared, one column can hold a brief description of these statistics (e.g., "Mean price for cement in 1997"), one can hold the number of firms used to generate this number, and the next 3 can hold your concentration ratios. Demonstrating non-disclosure this way is the recommended method since it makes the demonstration easy for you to organize and for your RDC administrator to follow.

Regression Output
A recent paper by Merrell and Reznek analyzes the disclosure risks from regression output.[1] One way to summarize the author's conclusions – which have been codified in the handbook – is that regressions are just a special case of tabular output. Regressions that involve only continuous variables can be thought of as a table with only one "cell."[2] In other words, the disclosure requirement for such a regression is to show, for the regression as a whole, that:

1. At least XX firms are used in the regression and
2. The YYY-firm concentration of TVS was no more than ZZZ percent.

(<u>**This rule is confidential…your administrator will verbally tell you X, Y, & Z**</u>.)

The exception, discussed in the handbook, is when you have dummy variables (independent or dependent) or fixed effects in your regression, and *you report the coefficients on these dummy variables.* In this case, the dummy coefficients represent conditional means. Therefore, the firms within each value of the included dummy variables (0/1) are the relevant "cells." See the RDC handbook or Merrel and Reznek's paper for further details.

Note that if you simply control for fixed effects but don't report their coefficients (for example state fixed effects) you need only show, as before, that the regression as a whole meets the two non-disclosure requirements.

<u>3. Providing a Benefit to the Bureau</u>
The final requirement for disclosure of research output is to provide the Census Bureau with a draft of your Post Project Certification (PPC). Your Post Project Certification, along with your project proposal, should include a statement concerning how your study might benefit the Census Bureau. At this point you need to include a discussion of how your study has actually benefited the Census. Here are a few guidelines for presenting to the Census, the benefits of your study.

First, the benefits must be substantive. By this I mean that merely stating (or restating) your benefit is insufficient for this purpose. Instead, you must now actually discuss and provide that benefit. For example, if you are doing some kind of data matching and analysis as one of your benefits, you should now actually provide that analysis. If you have brought in new data that would be useful to the Bureau, give the location and description of the data, etc. In the end, the document you produce may resemble a short

[1] Merrell, David R. and Arnold P. Reznek. "On Disclosure Protection for Non-Traditional Statistical Outputs." CES, US Census Bureau. Mimeo, July 1992.

[2] After all, a regression coefficient is just a normalized covariance statistic computed with all of the data.

research paper, including tables and figures. Perhaps you can think of it as a working paper on whatever technical subject that your benefit provides.

APPENDIX

There are many ways for you to generate non-disclosure statistics for your RDC administrator. Below are two available methods, one uses STATA and one uses SAS. Also, please note that since the concentration ratios are sensitive, whenever a reference is made to one it is replaced with a Y (e.g., CRY instead of CR987). When running these programs, simply replace Y with the number you are told by your RDC administrator.

STATA options for RDC Disclosure Requirements

If you typically use STATA, then there is an available program called “esum.do” that your RDC administrator should be able to provide you with. It should generate non-disclosure statistics for the variables you are trying to clear. In order to use esum, simply run “esum.do” after you’ve loaded your data. Then specify which variables you would like to disclose. Below are the instructions that come with esum detailing proper syntax and use. It should be noted that if you are trying to clear a large number of variables (say more than 10), esum typically works better if you make seperate requests (e.g., if you’re trying to generate non-disclosure statistics for 25 variables, use esum on the first 10, then the second 10, and finally the remaining 5). It should also be noted that if a syntactical error is made when asking esum to generate output, you may need to drop several generated variables before attempting to run esum again.

```
/*  esum was a program written to help rdc researchers pass disclosure, it
has a somewhat involved syntax, but in the end should greatly easy your ability
to quickly generate summary statistics for variables and datasets. Here's an
explanation of the syntax plus an example or two:

General Syntax:

esum <Variable List>, sType(<type>) other options

Options include:

sType - sType is asking you what sort of summary you would like.  At the moment
the only working options for sType are "mean" and "meannoz", where specifying
"mean" will generate the means based on non-missing observations for every
variable in <Variable List>.  "meannoz" however will generate the means based
on non-missing, non-zero observations for every variable in <Variable List>.

firmident - firmident allows you to specify the id variable.  For most CES
data,
the firm id lurks in the variable called "id" or "ID" however it might have
been
renamed during the course of being shuffled around and translated from one
format to another, hence you can specify the name here.

tvsident - tvsident is a tool for generating concentration ratios for your
variables in varlist.  If you choose to use it you must provide some total
value variable to accompany each variable in varlist.  Additionally, you
*must* specify firmident in order for tvsident to work.

showall - specifying "no" will suppress output.  This is a superior choice to
using "quietly esum" since it will increase the speed at which esum works.
```

```
Examples:

esum WaterUsage92 WaterUsage97, sType("mean")


                             Enhanced Summary Version 1
                             Displaying: Variable Means

     Variable   Number of      Aggregate       Mean        Standard
       Name       Obs            Value                     Deviation
-----------------------------------------------------------------------------
WaterUsage92  |     71       2.93025e+10      4.13e+08      8.3e+07
WaterUsage97  |     43       1.93872e+10      4.51e+08      7.9e+07


esum WaterUsage92 WaterUsage97, sType("mean") firmident("id")


                             Enhanced Summary Version 1
                             Displaying: Variable Means

     Variable   Number of      Aggregate       Mean        Standard   # of
       Name       Obs            Value                     Deviation  Firms
-----------------------------------------------------------------------------
WaterUsage92  |     71       2.93025e+10      4.13e+08      8.3e+07   11
WaterUsage97  |     43       1.93872e+10      4.51e+08      7.9e+07    7

esum TotalProducts BannanaShipments MonkeyBrains, sType("meannoz")
firmident("id") tvsident("tvs TVBannanas TVMonkeyBrains")


                             Enhanced Summary Version 1
   Displaying: Variable Means, not including observations with zeros

     Variable   Number of      Aggregate       Mean        Standard   # of
       Name       Obs            Value                     Deviation  Firms
-----------------------------------------------------------------------------
TotalProducts |     50             5234        104.68         8.238    35
BannanaShipme |     47             4502         95.79        12.543    34
MonkeyBrains  |     13              732         56.31         4.194     4


                             Concentration Ratings

resum2[3,3]
                     CRY        CRYY       CRYYY
TotalProducts     .44115612  .72435583  .90838208
BannanaShipments  .49241651  .76003673  .93341422
MonkeyBrains      .73523511          1          1
```

-Written by Philip Doctor
*/

SAS Code for RDC Disclosure Requirements

This section describes a SAS macro program that will produce LRD disclosure statistics on whatever tables you have produced in SAS. It was written originally by Dave Merrell and edited by me.

Aside: What is a SAS Macro?

If you’re not familiar with SAS macros, it’s analogous to a “subroutine” or “procedure” in other computer languages. It’s a set of SAS commands that can be called upon several times in a single SAS program with a single command. Like a procedure, it can be flexible: it can take user-supplied inputs that customize the use of those SAS commands. For example, you might want to be able to apply the macro to several different datasets, in which case you might write a macro that took the name of the dataset as an input, and applied all the commands to whatever dataset the user supplied.

This Macro

This is the macro. Note that at the top of the macro is text that describes how to use it. Below I also give examples of how it is used.

```
%MACRO extract(infile,outfile,byvars,wtvar,ifcond);
%* This macro was created originally by Dave Merrill, modified by Ethan Lewis.

   This macro generates disclosure analysis variables, specifically firm counts
   and shipment contentration ratios, for user-defined data tables.   It prints
   them in the SAS lst file and stores them in a user-named output dataset.


   INPUTS TO THIS MACRO:
   The user defines the table on which she desires to have disclosure statistics
   with three of the above variables:

   "infile" -- the name of the raw data source for creating the table.  Each ob-
              servation in this data source is an establishment.  (For example,
              the raw ASM data is an acceptible source.)
   "byvars" -- a list of the by variables, separated by spaces, which generate
              the cells of the table
   "ifcond" -- a logical condition that identifies the observations in
              the "infile" to be used in computing the table.  Note that
              you dont need to put the word "if" -- just the condition!

   The last two are optional -- they can be left blank (eg, if you want to get
   disclosure statistics for the dataset as a whole).

   The disclosure statistics are printed to the SAS .lst file, but they are also
   stored in the user-named "outfile" (not optional).

   Finally, the user should supply to the "wtvar" the asm weighting variable
   wt, if she is using ASM weights in her tabulations.


   VARIABLES IN THE "INFILE"
   The "infile" should contain (at least) the following two variables:
   - tvs (total value of shipments)
   - id (firm identification number)

   In addition, you may need to include the following variables:
   - If you are using the ASM weights in your computation, the wt (weighting)
     variable should be in "infile."  (Also dont forget to set wt for
     "wtvar." [See above.])
   - If your table has any "by" categories, these variables should be
     included in "infile."
   - If you are filtering the data with an "ifcond," the variables in your
     condition should be included.


   VARIABLES IN THE "OUTFILE"
   The output datset will contain the following variables:
   - The user-supplied by-group variables ("byvars"), if any.
   - crY = the Y-firm concentration ratio for this cell of the table
   - firmct = the number of firms in this cell of the table
   - disc = "D" if this cell of the table cannot be disclosed because it
     would violate confidentiality.  Otherwise it is blank.


   MACRO OUTLINE:
   The outline of this program is as follows:
      1. Create a dataset with Firm ID variable that identifies establishments
         that are part of the same firm.
      2. Compute the Y-firm (CONFIDENTIAL!!) concentration of shipments.
      3. Count the Number of Firms in the Cell.
      4. Merge the Y Disclosure statistics, summarize and print out.

   ;


   %*----------------------------------------------------------------------;
   %MACRO dim(list);
   %* Returns the number of elements in "list" (which must be separated by
      spaces -- easily reprogrammed.)  This macro-within-a-macro is used here
      to calculate the number of elements in "byvars" "ifcond" and "wtvar" from
      above.;

      %local i item;
      %let i=1;
      %let item=%scan(&list,&i,%str( ));
      %do %while(&item ne);
        %let i=%eval(&i+1);
        %let item=%scan(&list,&i,%str( ));
      %end;
      %let i=%eval(&i-1);
      &i

   %MEND dim;
   %*----------------------------------------------------------------------;


   %let nbyvars=%dim(&byvars);
   %let byst=;
   %let wtst=;
   %if &nbyvars>0 %then %let byst=by &byvars;
   %if "&wtvar"="wt" %then %let wtst=weight wt;


   %* 1. Create a dataset with Firm ID variable that identifies establishments
         that are part of the same firm;
   data __temp;
     set &infile ;
     %if "&ifcond" ne "" %then %do; if (&ifcond); %end;

     id1=substr(id,1,1);
     id6=substr(id,1,6);
     idstripd=id6||'0000';
     if (id1=0) then firmid=id;
     if (id1>0) then firmid=idstripd;
     drop id1 id6 idstripd;

   proc sort;
     by &byvars firmid;


   %* 2. Compute the Y-firm (CONFIDENTIAL!!) concentration of shipments;

      %* a. Compute NUMERATOR: The sum of shipments (TVS) at the Y
            biggest-shipping firms within each by-category ;
         %* i. Compute Shipments at each firm in the by-category;

            proc summary data=__temp;
             by &byvars firmid;
            output out=numfirm
             sum(tvs)=tvsftot;
            run;
         %* ii. Rank these Firms Based on Total Firm Sales;
            proc rank data=numfirm descending out=tvsfirm;
             &byst;
             var tvsftot;
             ranks tvsrank;
            run;

         %* iii. calculate sum of top Y-firms shipments;
         data tvsfirmY;
           set tvsfirm;
           if tvsrank<=Y then output;  %* top Y firms;

         proc summary;
          &byst;
         var tvsftot;
         output out=tvsfirmY sum(tvsftot)=tvstotY;
         run;


      %* b. Compute DENOMINATOR: Shipments in Each By-Category as a Whole;
      %* Notice that the denominator is weighted sum, while the numerator
         is not.  This follows required disclosure rules (see RDC handbook);
      proc summary data=__temp;
       &wtst;
       &byst;
      output out=numfirmY
       sum(tvs)=tvstot;
      run;

      %* c. Compute the ratio;
      data crY;
        merge tvsfirmY numfirmY;
         &byst;

        crY = tvstotY/tvstot;


   %* 3. Count the Number of Firms in the Cell;
   proc freq data=__temp;
     tables
     %if &nbyvars>0 %then %do;
        %scan(&byvars,1)*
        %if &nbyvars>1 %then %do;
           %do i=Y %to &nbyvars;
              %scan(&byvars,&i)*
           %end;
        %end;
     %end;
        firmid / noprint out = firmdist;
   run;

   %if &nbyvars>0 %then %do;
      proc sort data=firmdist;
       by &byvars;
      run;
   %end;

   proc summary data=firmdist;
    &byst;
    output out=firmct
     n(percent)=frmct;
   run;


   %* 4. Merge the Y Disclosure statistics, summarize and print out;
   data &outfile;
     merge crY firmct;
      &byst;

     /* Census Disclosure Review Policies specify NN and MM below */
```

```
        if frmct<NN or crY>.MM then disc='D';

        keep &byvars crY frmct disc;

      proc print;
      title "Disclosure Statistics For This Table";
      format crY 6.3;
      var &byvars crY frmct disc;
      run;


      %* Clean up memory;
      proc datasets;
        delete __temp crY firmct numfirm numfirmY tvsfirm tvsfirmY;

   %MEND extract;
   *------------------------;
```

So, for example let’s suppose you are working with 1992 Census/ASM data, which you extract into a data file “rawdata.”:

```
Data rawdata;
  Set cm.cmf0092 (keep=id tvs reg et wt dind);
```

…where the “reg” variable is major census region (1-4); “et” is 1 for non-ASM establishments, 0 for ASM establishments, and “dind” is the SIC industry code. The following command:

```
%extract(rawdata,asm92,reg,wt,2010<=dind<=2099);
```

…will take the rawdata data and produce a dataset “asm92” giving for the firms in food industries (2010<=dind<=2099), by region, the Y-firm concentration of shipments and the number of firms. Since the “wt” variable option is used, it assumes you are just working with the 1992 ASM observations and weights the computation of these disclosure statistics appropriately.   This disclosure table would be the companion to, say, a table of mean capital stock by region in the food industry in 1992 that you wished to take out of the RDC. The disclosure table might look something like this:

```
proc print data=asm92;
run;

obs     reg      crY      frmct     disc
1       1        0.3421   253
2       2        0.6633   8         D
3       3        0.2222   885
4       4        0.9223   16        D
```

(NOTE: These data are entirely made up.) This shows that publishing your statistics from regions 1 and 3 is OK. However, computations involving food establishments in region 2 pose a disclosure risk because of the small number of food firms in that region. There are sufficient numbers of food firms in region 4, but statistics on region 4 food establishments also cannot be published because shipments are so highly concentrated there.

More examples…:

If you are computing unweighted statistics using the Census of Manufacturers, just leave out the “wt” part, as in:

```
%extract(rawdata,cm92,reg,,2010<=dind<=2099);
```

You can use more than one “by” variable, or even none! For example:

```
%extract(rawdata,asmcm92,reg et,,2010<=dind<=2099);
```

would compute disclosure statistics for ASM and non-ASM food establishments by region (unweighted). Thus, the output dataset “asmcm92” would be a table with 8 cells (4 regions x ASM/non-ASM).

Here’s an example of how to do it when you have “no” by variables (for example for disclosure of a regression that used all establishments in the food industry):

```
%extract(rawdata,cm92food,,,2010<=dind<=2099);
```

Finally, you can also leave out the “if” condition if you want to look at all of the observations in your dataset. A couple of examples:

```
%extract(rawdata,cm92reg,reg);
%extract(rawdata,cm92all);
```