### Record Linkage to External Data

*[The confidentiality analyses must be designed to provide reasonable assurance that the public-use data files will not allow identification of individual schools, (or other institutions), students, teachers, or other individuals when compared to publicly-available data collections.]* No publicly-available data collections identify either students or teachers by name, but the following two data collections identify schools by name. The NCES regularly publishes the CCD, a detailed public school listing, and the PSS, a detailed private school listing. Any potential identification of teachers and students will arise through the identification of their associated schools. Providing a reasonable degree of assurance that schools cannot be identified thus generally assures that teacher and student data also remain unidentifiable. Nevertheless, the confidentiality procedures included perturbing student- and teacher-level data.

A data sleuth who wants to identify the schools could take school data and search the CCD and PSS files for matching data by first identifying variables common in both files. The user then could compare a school record to each CCD school record, assigning each pair a probability of being a true match based on the common variables. The CCD record (or records) with the highest probability would best match the traits of the study school. A study school could then be identified from the school name and address attached to the CCD school record. The user could take a similar approach in identifying a study private school from the PSS school file.

A data user might find one or more schools on the CCD or PSS files with the highest probability match to a study school. As the study public-use data files do not contain any data that uniquely identify the school, the user would not know for certain whether any of the matched schools were the actual study schools. Duplicating how a user would match the study school file against the CCD and PSS school files can determine if the actual study school is among the schools identified as the top matches. Perturbing data on the study school record such that the study school can no longer be found among the top matches increases the confidentiality of the study school data.

[*Data disclosure definitions depend on a variety of factors, including the availability of public data resources, the effectiveness of matching strategies and software, and the effectiveness of previous confidentiality analyses on similar data sets. The operational definition of confidentiality comes from the NCES “rule of three” statistical matching guideline. The public-use data files can be considered confidential if the matching of study school records with CCD and PSS school records demonstrates the following criteria:*

* *The schools identified as the best match and the next highest match in the CCD file are not the study school in question;*
* *The schools identified as the best match and the next highest match in the PSS file are not the study school in question; and*
* *If the CCD or PSS files have between three and five schools tied as best matches, none of these schools are the STUDY school in question.*

*One of the problems with the external matching procedures for the IES DRB has been that different IES studies have used different approaches and procedures for generating the external matching routines. Various home-grown Euclidean distance programs have been used to identify matches in early studies, but there are some problems in terms of accuracy with this methodology. Beginning in 1995, many studies used AutoMatch software (Jaro 1989) that incorporated probabilistic record-linkage matching procedures that could better determine whether the survey record was a likely match with an external file record. The IES DRB prefers standardized software that allows them to better evaluate the validity of the procedures and the results. That is, using approved software that generates standardized results and reports enables the IES DRB to more quickly and confidently determine whether the matching procedures were properly implemented. The IES DRB also developed a software package, DRISK, to emulate AutoMatch and provide to all contractors. However, both DRISK and AutoMatch are older programs that cannot run on newer operating systems.*

*The software suggested to conduct the matching of study data to the CCD and PSS has been identified by the IES DRB. It is a probabilistic record linkage software package called Fine-Grained Records Integration and Linkage tool (FRIL). It is the preferred approach though others can be used if acceptable to NCES. FRIL is offered by the Centers for Disease Control. Brief information regarding to FRIL parameters and processing is provided below. The documentation is linked to the DAP-TS.]*

##### FRIL Parameters & Processing

[*Provide key processing information including the following*]

Create Matching Files

Compare Identifiers across Paired Records

Blocking Variables

Matching Variables

Conduct Probabilistic Matching Analysis

Masking Matched Records (e.g., deterministic swapping)

Rerun FRIL to make sure the change is sufficient.

##### FRIL Output

[*The output from FRIL can be put in an appendix with tables with information as shown below, for example. Briefly explain what those output are*.]

[Study name]/CCD Matches and Duplicates -

SCHOOLID=0001

TYPEREC NCESSCH MATCH RURAL SCHCOUNT GRDCOUNT PMAL PFEM CONFIDENCE