## Review of Common Data Elements for Potential Matching

The first step in the confidentiality analyses identifies the set of school attributes held in common by the IES Study/CCD and IES Study/PSS data sets. Post-secondary studies match with IPEDS data (mostly found in the Institutional Characteristic file). Logically, records can only be matched on variables the data sets share in common. There are two sets of variables that will be used to determine whether a school can be identified. The first set of variables will be the IDs, the weights and sampling variables. We must analyze these data individually, and together to determine whether these variables can be used to identify an actual state or other entity (district, MSA). If a state can be identified, then that information would be included with the background variables in the matching procedures or, in the case of IDs, masked. Prior to conducting the matching component of the disclosure analysis, the contractor should consult with the NCES staff to determine which variables should be retained, if possible, in the public-use file release. The variables requested for data inclusion by NCES may include several "risky variables" - such as school type, Census region and type of location - that are exact matching variables that we use to block-match schools with the CCD/PSS/IPEDS.

Some studies pull variables from the CCD/PSS/IPEDS and include them with the survey data and derived variables on the delivery files. These would be the riskiest of all variables since they would be guaranteed to be an exact match unless they were somehow first perturbed.

*Please refer to the section on probabilistic matching criteria in order to determine how to best utilize the blocking/matching variables to optimize the identification of disclosure risk schools.*

The variables described below are the common matching variables. They are not exhaustive as other variables not mentioned may also pose a disclosure risk via matching.

### Study ID

The study IDs should be randomized IDs not linked to any geographic or identifying location. Some studies may include identifying IDs. If so, they must be replaced for public-use dissemination.

### Geographic Identifiers

State is sometimes a key variable in studies – the sample was designed for a sufficient number of schools in all participating states. This allows for valid state statistical analyses. However *any reference to an actual state or district must be suppressed.* This determination is based on extensive disclosure analyses we had conducted on a number of k-12 studies. Having the geographic information about a state or district on the file renders the data from many states as disclosure risks. Suppression of various demographic data were evaluated, and the results were reviewed by NCES. Based on these analyses, NCES determined that the state school data cannot be disseminated publicly on a number of survey studies. Based upon previous analyses from various survey data, Region can be identifying.

In summary, all geographic information would have to be removed (city, zip code, state, LEAID, etc.).

## Weights and Sample Design

Weights are a concern if they can uniquely identify a school or geographic location (state, district). If weights are highly correlated to FTE (full-time equivalent number of teachers), or enrollment and anyone with the CCD (which is public-use) will have access to the FTE and enrollment distribution across schools within a state, which should be fairly unique state-to-state and would make small states vulnerable to disclosure risk. Knowing how NCES samples schools, which is a matter of public record, can provide some clues to data sleuths, particularly if any geographic information is available in the data. Added information about grade span, enrollment, minority breakouts, charter status, etc. will make identification even easier. Knowing the weights alone with nothing else perturbed should allow a fairly unsophisticated user to possibly guess at states. In summary, weights, in conjunction with other existing demographic information on the file, can potentially be used for school identification. Since the weights are correlated to the number of teachers in the school, then suppressing or collapsing this variable would not be helpful since the weights can be used to reproduce them.

## Demographic Variables

The primary issue concerning the release of a public-use file is that many of the data elements in the file can be used to positively match against the CCD and PSS and IPEDS using probabilistic record linkage software.

The variables that can be used for matching can be split into three categories: (1) identifying variables – requiring suppression; (2) exact matching variables – may require suppression or collapsing; and, (3) likely inexact matching variables with some designated tolerance that may require suppression or collapsing:

Not all variables from the survey data and/or matching public files are direct matching variables. Some may be derived to better match. For example, we can use and/or derive the percent gender, percent minority, percent free lunch, in the school etc. (from aggregated student data or school-based variables) and match these variables with comparable variables found in the CCD, PSS and IPEDS. Historically these derived variables have been fairly predictive of school level data found in the public data bases. If we do conduct the matching analysis and identify matches (a school is considered a disclosure risk under the "Rule of 3"), then we would need to further coarsen or delete variables in order to eliminate the disclosure risk.

## Typical Identifying Variables (often for immediate suppression):

- Variables drawn from the publicly available files (CCD/PSS/IPEDS)
- (CCD/PSS/IPEDS ID's) – serves as current school ID for public and private schools
- All state references (FIP code, etc.)
- Zip code

- Charter school identifier
- School control number
- Principal control number
- Census region, based on ANSI state code
- School locale code (urbancentric 12 categories)

### Exact Matching Variables:

- Grade Span – all grades found in the school
- Flag indicating enrollment of American Indian students
- Three-category school level (elementary/secondary/combined)
- Four-category school level (primary/middle/high/combined)
- Sector (public, private)
- Affiliation
- Year round school
- Collapsed school locale code (urbancentric – collapsed 4 categories)
- Flag indicating a school-wide magnet program

### Inexact Matching Variables:

- Total K-12 and ungraded enrollment in school
- Estimated number of full-time equivalent teachers in the school
- Percentage of enrolled students approved for the NSLP at school
- Estimated percentage of students who are American Indian/Alaskan Native (not of Hispanic or Latino origin)
- Estimated percentage of students who are Asian (not of Hispanic or Latino origin)
- Estimated percentage of students who are Black (not of Hispanic or Latino origin)
- Estimated percentage of students who are female.
- Estimated percentage of students who are of Hispanic or Latino origin
- Estimated percentage of students who are male.
- Estimated percentage of students who are two or more races (not of Hispanic or Latino origin)
- Estimated percentage of students in school who are non-White
- Estimated percentage of students who are White (not of Hispanic or Latino origin)
- Estimated number of students per FTE teacher in the school
- Percentage of enrolled students who are LEP
- Percentage of enrolled students with an IEP
- Three-level private school typology
- School days (number of days in the school year)
- Online courses – number

- Flag indicating whether school has students enrolled in the International Baccalaureate Diploma Programme
- Enrollment (school)
- Magnet program
- Title I services

For a complete review of all potential matching variables, please refer to the following sources:

You can download IPEDS data at:
https://nces.ed.gov/ipeds/datacenter/login.aspx?gotoReportId=7

You can download the CCD files at: https://nces.ed.gov/ccd/ccddata.asp

You can download the PSS files at: https://nces.ed.gov/surveys/pss/pssdata.asp