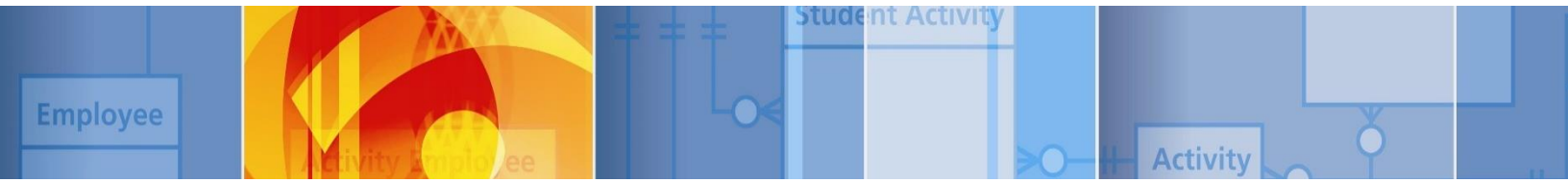


THE OPTIMAL REFERENCE GUIDE

De-Mystifying De-Identification

Extraordinary insight™ into today's education information topics

By Glynn D. Ligon, Ph.D., ESP Solutions Group, Inc.

Intended Audience: If you manage or use an education database and are concerned about confidentiality, read this paper. There's no need for extreme measures that remove too much personal information from all of our databases. Polititech decisions can support official reporting, research, FOI requests, and FERPA/HIPAA.

ESP Solutions Group

A Primer on De-Identification Jargon

This paper doesn't ease the reader into the milieu of de-identification. A brief primer may help establish a baseline understanding of terms.

- Data Governance: An education agency's policies and processes for overseeing the collection, storing, accessing, and reporting of its data
- De-Identification: Any method used to remove or obscure data elements or values associated with a single individual
    - De-identification is important because individuals or their parents want to preserve the confidentiality of information about them.
- FERPA: The Family Educational Rights and Privacy Act (1974) that is the foundation for protecting education records
- HIPAA: The Health Insurance Portability and Accountability Act (1996) that protects health records
- Longitudinal Data System (LDS): Data warehouse, database, data store, data mart, dashboard, portal—any and all combinations of these systems that collect, store, and report data across years (With an S in front, the SLDS designates a statewide system. With P20W added, the P20W SLDS designates a multiagency statewide system.)
- Polititech: The merging of politics and technology to create data governance policy and processes
    - Polititech is important because data governance is the solution to resolving the paradox of protecting confidential personal data while maintaining the integrity of the contents of the longitudinal data system.

De-Mystifying De-Identification

Introduction

We have two goals when de-identifying the records within a database. Simultaneously, we intend to:

1. Remove the personally identifiable characteristics of individuals, and
2. Retain the integrity of the records for analysis and reporting.

As soon as we set out to establish the business rules for de-identifying individual records, the paradox of these goals becomes apparent.

Consider Isaac Newton's Third Law of Motion: For every action, there is an equal and opposite reaction. (Mathematical Principles of Natural Philosophy, 1687)

Let's restate this simply for databases. Every action we take to de-identify a data element in our records creates an equal and opposite reaction against our capability to analyze and report from our database. Unfortunately, Mr. Newton lived long before we discovered that the reactions within databases are exponential. Therefore, deleting a single data element from a database can in reality disable untold combinations and permutations of relationships and causalities available to a researcher to explore. In other words, we should be so lucky, Mr. Newton, to suffer only a single equal and opposite reaction for each data element we tamper with in our database.

The message in this paper about de-identification is:

- Do it with full knowledge of the degradation of the research and analytic value of the database;
- Do it at the least disturbed level allowable; or
- Better yet, don't do it.

If you don't do it, then what's the alternative? There're four steps to protecting the identities of individuals and still allowing their personal information to reside intact in a database.

1. Secure the database from unintended access.

2. Authorize the users for approved purposes.

3. Authenticate the users upon accessing the data.

4. Mask the data in any small reported cells.

Whatever your choice, a data governance policy should protect the personally identifiable data within your agency's databases without question. However, an enlightened data governance policy will also enable access to identified data for authorized purposes by authenticated individuals. So, this paper's guidance is to de-identify a database for research purposes if necessary, but to rely as often as possible on vetting the researcher for access to the full data. Then police the masking of published results to hide small cells, and never forget to apply rules that require cells to be reliably large as well.

Did you notice I tossed in the reliability criterion? Too often we forget that if our reporting followed protocols for publishing statistically reliable numbers, that those numbers would always be large enough to protect the identities of the individuals in the reported cells. Thus, our data governance should not overlook establishing and enforcing reliability rules for reporting.

Fortunately, an education agency has the option of having more than one database. The "don't do it" admonition doesn't really apply unless an agency is going to restrict itself to a single data store.

What's the bottom line for an education agency? Develop and publish a data governance policy that specifies three processes.

1. Analysis and Official Reporting from the Longitudinal Data System
2. Research and Evaluation from De-Identified Data
3. Reported Data with De-Identified Small Cells

Politimetrics is known as decisions made through a combination of science (psychometrics) and policy (politics). Examples are where to set the score for proficiency, how many credits to require for graduation, and what score qualifies a student to enroll in a special program. Neither the pure psychometricians nor the pure politicians should make these decisions independent of the data and wisdom of the other. **Polititech** is data governance. Managing individual identities into and out of databases is quintessential polititech. Designing a database model to fit the political (e.g., FERPA and HIPAA) mandates of governing law, the compliance reporting rules of enabling legislation, and the analytical requirements of researchers is…challenging. There should be no surprise that more than one database is required. Likewise, there should be complete agreement that a single data governance policy overseeing everything is essential.

De-Identification

FERPA and HIPAA do not restrict an agency from collecting and storing personally identifiable data on its students, employees, and those it certifies; however, there are some restrictions on how those data persist and are shared when the individuals terminate their relationship with the agency. This paper is not about the laws. This is about polititech—how the laws intersect with technology. So our focus will be on data governance issues.

Polititech requires that our data governance policy guide the design of information systems to support three processes.

1. Analysis and Official Reporting from the Longitudinal Data System
   The agency must have a longitudinal data system with unmodified records for official purposes and reporting. Providing access to authorized experts with purposes consistent with the data governance policy serves the goals of the agency. These experts would include agency analysts as well as approved external researchers.

2. Research and Evaluation from De-Identified Data
   The agency must have a de-identified database to provide with confidence to external researchers. Providing a readily available de-identified database for external analysts is a practical and cost-efficient process for an agency to respond to freedom of information requests as well as academic proposals. External researchers would include anyone with a legitimate request meeting the data governance policy's guidelines or a freedom of information request's criteria.

3. Reported Data with De-Identified Small Cells
   For publications deriving from any source, the data governance policy must specify acceptable processes for de-identifying small numbers in reports that might reveal personally identifiable information.

These processes do not need to be supported by the same database.

These processes do not necessarily serve the same users.

With those parameters established, discussing de-identification is much more focused. This paper is organized into three sections, one for each of the processes identified.

In each section, these issues will be discussed.

- Users: Who are the people accessing the data?
- Questions: What are the questions these users ask of the data?
- De-Identification Methodology: What de-identification processes are available?

The Data Governance Policy

De-identification is one issue within an education agency's Data Governance Policy. The polititech that drives it comes mainly from FERPA, state laws, and local policies derived from those. Generally, these are among the processes that should be incorporated within the agency's Data Governance Policy.

- MOUs with other agencies sharing data for analytic and research purposes
- Three levels of identification/de-identification/re-identification
    - Identified for official statistics, reporting, compliance, research, and evaluation
    - De-identified for FOI requests, external research, and public reporting
    - Public reports
- Metadata dictionary for defining data elements, transformation rules, ETL processes, and database tables
- Business rules and data element definition standards for vendors
- National and other standards followed by the agency

Section 1: Analysis and Official Reporting from the Longitudinal Data System

> History Lesson: FERPA was passed in 1974 primarily to ensure parents' rights to access and control access to their children's records. HIPAA was not passed until 1996 partly to protect the confidentiality of patients' records. When FERPA emerged, most student records were on paper. The Federal Migrant Student Record Transfer System began collecting individual records in 1969. Local education agencies have collected automated individual records in their student information systems since those first emerged in the 1970's. Florida and Texas were the first states with mass collections of individual records in the 1980's. Before the No Child Left Behind Act of 2001, the collection of individual student records by state education agencies was the exception, not the rule as it is today. The practical reasons for education agencies to collect individual records instead of aggregate statistics are efficiency and data quality.

This is the education agency's core data warehouse. An education agency's longitudinal data system (LDS) must have unmodified records for calculating complete official statistics and reporting. Every mandated detail must be maintained in the database for reporting and audit purposes. If data elements are de-identified, then the burden falls back to a prior level of reporting for audit purposes.

Providing access to authorized experts with purposes consistent with the data governance policy serves the goals of the agency. These experts would include agency analysts as well as approved external researchers. Identity management systems can control each person's authority to access specific areas of the database and the actions each person can perform. Each person is authenticated upon sign on and authorized as to the permissions assigned.

A key component of the data governance of the LDS is the agency's metadata dictionary. This essential guide contains and manages the definitions, business rules, transformation formulas, table formats, ownerships, and other relationships for all collections, repositories, and outputs (i.e., reports, publications, and other media coming from the LDS or any of its related data marts or dashboards).

Longitudinal Data System		
Users	Internal	Program Officers, IT Staff, Agency Officials
	External	Approved Researchers, Contractors
Questions	Internal	• What are our agency's official statistics? • What students meet early warning criteria? • What schools met annual accountability objectives?
	External	• Did X Reading Program outperform Y Reading Program for individual subgroups in district Z?

		• What was the impact of changes in graduation requirement policies for individual subgroups in District Z?
De-Identification Methodology		None

Section 2: Research and Evaluation from De-Identified Data

The agency must have a de-identified database to provide with confidence to external researchers. Providing a readily available de-identified database for external analysts is a practical and cost-efficient process for an agency to respond to freedom of information requests as well as academic proposals. External researchers would include anyone with a legitimate request meeting the data governance policy's guidelines or a freedom of information request's criteria.

De-Identified Database		
Users	Internal	Research and Evaluation Staff
	External	Researchers, FOI Requestors
Questions	Internal	• Have statewide performance level trends changed?
	External	• Did X Reading Program outperform Y Reading Program statewide? • What was the impact of changes in graduation requirement policies statewide? • Has enrollment in charter schools changed?

De-Identification Methodology		• Safe Harbor
		• Expert Determination
		  ○ Anonymization
		  ○ Blurring
		  ○ Record Code
		  ○ Suppression
		  ○ Any Other

HIPPA has made it clear that there are two methods to achieve de-identification in accordance with their privacy rule. The two methods contrast greatly in their specificity. The first is to have an expert determine a method that works and certify it. What constitutes an expert and what criteria that expert uses are entirely up to the agency. On the other hand, the second method, safe harbor, is to suppress

in the records 18 specified data elements for the individual or the individual's relatives, employers, or household members; and to certify that the agency has no actual knowledge that the information could be used alone or in combination with other information to identify the individual. Exhibit A is the full description of HIPAA's methods and the data elements they define to be removed.

Under expert determination, what methods might be acceptable for education agencies? The Privacy Technical Assistance Center has defined several methods in its brief, "An Overview of Basic Terms."

Methods have been defined, precisely and poorly, by multiple authors over the years. So much so that citing them selectively would over emphasize their completeness and official stature. So this paper will summarize the terms and definitions in a manner not pretending to be comprehensive, but merely introductory. The contribution made here will be to attempt to differentiate the terms and methods from each other; whereas, in the literature to date, some have been loosely applied.

- Anonymization
- Categorization of Continuous Variables
- Substituting Individual Values for Group Averages
- Controlled Rounding
- Combining Cells
- Suppression
- Top/Bottom Coding
- Transformation Algorithm
- Data Swapping
- Random Misclassification
- Record Code Substitution (Tokenization)
- Redaction
- Encryption

Noticeably absent from this list are some commonly referenced terms (e.g., masking, perturbation, noise, disclosure limitation, and disclosure avoidance). However, these terms refer to generalized categories of techniques inclusive of the ones defined above, not methods themselves.

These include the following.

- Masking (reconfiguring data)
    - Categorization of Continuous Variables
    - Substituting Individual Values for Group Averages
    - Controlled Rounding
    - Combining Cells
    - Top/Bottom Coding
- Perturbation/Noise (changing data)
    - Data Swapping
    - Transformation Algorithm
    - Random Misclassification
- Disclosure Limitation (holding back data)
    - Anonymization

    - Suppression
    - Redaction
- Disclosure Avoidance (denying data)
    - Disapproval of Requests

Another somewhat confusing concept in the discussion of de-identification is the distinction between:

- Treatments to data in fields within a database and
- Treatments to reported data in published tables.

The best way to conceptualize this might be that all de-identification techniques apply to databases because all data from their raw state to their derived statistics in tables are stored in databases. Therefore, the need to de-identify the same data represented in published tables in their representation in an underlying database exists. Thus, all the de-identifying techniques are mentioned in this section, but only those that are particularly appropriate for published tables are included in Section 3.

Just to restate, this isn't a user manual on how to perform these functions. So, what follows is an overview of what each technique is and how it is appropriately applied.

Masking and **blurring** are terms too often thrown around loosely as if they really refer to specific techniques. Instead, masking is a category of methods for reconfiguring data. The purpose of masking is simply to minimize the possibility that anyone could reconstitute the identity of an individual in a reconfigured group. These techniques apply more to group measures of central tendency than to individual's values. Therefore, they would modify aggregate statistics within a database more often than a field within an individual's record. However, as seen below, because one of the techniques itself is substituting individual values for group values, these can be applied to fields for individual records.

In Section 3 examples of some of these techniques, which are used in public reporting, are presented.

These very brief definitions help differentiate these techniques from each other.

    - Categorization of Continuous Variables
        - Converting a continuous variable into categories can prevent someone from recovering a cell's/field's value using a total and other cell/field values.
    - Substituting Individual Values for Group Averages
        - With only the group average, recovering the precise value for an individual within a group is less likely.
    - Controlled Rounding
        - Rounding individual values that are represented as decimals can prevent someone from recalculating a cell's/field's value using a total and other cell/field values; or recalculating an individual value within a cell/field.
    - Combining Cells
        - Combining two or more small cells/fields to create a larger group that meets the minimum size for reporting effectively achieves the confidentiality mandate.
    - Top/Bottom Coding
        - Creating a range of values at the top or bottom that includes a large number of individuals and reporting ranges throughout prevents identification of individuals when few appear at the very top or bottom of the range.

- Perturbation/Noise (changing data)
    - Data Swapping
        - Values are exchanged between individuals.
    - Transformation Algorithm
        - A formula is used to create sample data or to rearrange data.
    - Random Misclassification
        - Individuals are randomly moved among classes/groups.
- Disclosure Limitation (holding back data)
    - Anonymization
        - An individual's personally identifiable information is removed.
            - Record Coding/Tokenization
                - A random, identifier with no intrinsic meaning is substituted for an official one to enable longitudinal or cross file linking.
                    - Re-Identification
                        - The original identifier is reinstated; however, this reconstitutes the record as personally identifiable.
            - Safe Harbor
                - Measures are followed to meet HIPAA's criteria (see Exhibit A).
    - Suppression
        - Data are removed from a record.
    - Redaction
        - Data are edited from the results of an analysis or report.
- Disclosure Avoidance
    - Denial of Requests
        - A decision is made not to respond positively to a request for data.

Section3: Reported Data with De-Identified Small Cells

For publications deriving from any source, the data governance policy must specify acceptable processes for de-identifying small numbers in reports that might reveal personally identifiable information.

De-Identifying Small Cells		
Users	Internal	All Staff
	External	Researchers, All External Data Users
Questions	Internal	• Have statewide performance level trends changed?
	External	• Did X Reading Program outperform Y Reading Program statewide? • What was the impact of changes in graduation requirement policies statewide? • Has enrollment in charter schools changed?
De-Identification Methodology		• Cell Suppression • Sample • Limit Detail • Top/Bottom Coding • All Others

What techniques are available for de-identifying small cells without allowing for recalculation or excessive obfuscation? When are there too few individuals in a subgroup to allow disaggregating that will not reveal personally identifiable information for those individuals? Every education agency's data governance policies and processes must clearly describe the answer for these questions. Back in 2001, the intent in NCLB was to remove the possibility that this accountability system would require states to violate the established federal protection of student privacy as mandated under section 444 (b) of the General Education Provisions Act (Family Educational Rights and Privacy Act (FERPA) of 1974). Thus, if a subgroup is so small that publishing the percent proficient would reveal how an individual student scored, the state is not required to disaggregate the subgroup, and the school is neither responsible for reporting on this subgroup, nor responsible for this subgroup's meeting the annual objectives.

The majority of the content in this section is drawn from three prior papers.

- Ligon, G. D., Clements, B. S. (2008). Revisions to FERPA Guidance. ESP Solutions Group, Inc.
- Ligon, G. D. (1998). Small Cells and Their Cons *(Confidentiality Issues)*: NCES Summer Data Conference.
- Ligon, G. D., Clements, B. S., & Paredes, V. (2000). *Why a Small n is surrounded by Confidentiality: Ensuring Confidentiality and Reliability in Microdatabases and Summary Tables.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

This discussion makes several assumptions that should guide the polititech of an agency's data governance policies and processes.

- The Family Educational Rights and Privacy Act (FERPA) is the primary federal mandate to be followed.
- The values for subgroups with too few individuals to protect the identities of those individuals should be de-identified in all public reports.
- De-identified values should not be recoverable through calculations using other published statistics, e.g., the values of other subgroups or values published in separate documents.
- The existence of a de-identified subgroup should not require the de-identification of other sufficiently large subgroups to satisfy the previous assumption.
- The same minimum number of individuals should apply to all schools and districts, and the state in the calculation of accountability determinations. (This is an equity issue and a control to avoid manipulation of the rules to benefit individual schools or districts.)

Data collected by governmental agencies must remain confidential in order to protect the privacy of individuals. For the Census Bureau, that information may be related to geographic region, such that information reported for a sparsely populated area can easily be tracked to the few individuals who live in that area. For the Internal Revenue Service, it may be related to income, in that certain income levels are only attained by a few individuals. For educators, it can be information about test scores, disabilities, or socioeconomic status that must be reported in a way that does not reveal information about individual students or employees.

If, for instance, there are two Asian students in the fourth grade of a school and the percent proficient for Asian fourth graders is 50%, the parents of each of those students, knowing their own child's proficiency level, can easily figure the other child's. Alternatively, if there are 100 Hispanic students in the fourth grade, and the percent proficient for Hispanic fourth graders is 100%, then it can be easily determined that each Hispanic student scored at the proficient level. However, important information on subgroups must be reported. Certainly the taxpayers of a school district want to know if students of one gender or ethnicity lag behind others in test achievement. The task becomes finding a way to report enough information while still protecting the privacy of individuals.

Evans, Zayatz, and Slanta (1996) address data confidentiality issues faced by the Bureau of the Census. As in education, "The disclosure limitation problem is to prevent data users from being able to recover any respondent's reported values using values appearing in the published tables" (Evans, et al., 1996). They note that cell de-identification is a choice, but while de-identifying individual cells can be done relatively easily, de-identifying those cells in associated documents can be overwhelming. In this case, if the number of subjects in any cell is fewer than a certain number, that cell is de-identified from any data presented to the public. While this is fairly simple, it becomes more complicated because those cells may be carried over onto other data tables, and must be de-identified there, as well. In addition, revealing any cells which could lead to the exposure of the values in a small cell must also be de-identified. It is conceivable that this situation could lead to the loss of information for all subgroups. As noted earlier, it is unacceptable in an accountability system to lose information unnecessarily.

Adding noise to data tables is suggested as an alternative by Evans, et al. (1996). This means multiplying the data from each establishment by a noise factor before tabulating the data. Over all establishments, the number of positive (>1) and negative (<1) multipliers would be equal, so that they would cancel each other out in the end. Cells which appear in more than one data table would carry the same value to all tables. Zayatz, Moore, and Evans point out, however, that if the number in a cell is too small (1 or 2) it can still be possible to discern a unique contributing entity. Winkler (1997) observes that introducing enough noise to prevent re-identification of records may also make the files analytically invalid.

Moore (1996) identifies three other methods used by the Census Bureau. They are (1) release of data for only a sample of the population, (2) limitation of detail, and (3) top/bottom-coding. The first is not practical for the field of education. Information released must be based upon all students in all schools. The second, limitation of detail, is practical and useful in education. The Bureau restricts release of information which would be restricted to a subgroup less than 100,000. Educators use a much smaller limit, but as mentioned above they do, in fact, restrict release of information about subgroups which do not meet a certain size. The third method, top/bottom-coding, is very appropriate to the field of education. The Census Bureau limits reported levels of income because they might identify individuals. So incomes above a certain level, which might lead to identification of individuals, are reported as "over $100,000."

Numbers of students in a subgroup can be reported in a similar way. The following is an example of a way to report information about the percent of students who passed an assessment with a score of "proficient" using limitation of detail. See Table 3.

Table 3: Limitation of Detail Using Ranges for Number of Students						
	Total Students	**African American**	**Hispanic**	**White**	**Asian**	**American Indian**
% Proficient or Above	77.39	90	85	70	80	*
Number of Students in Group	115	5 to 15	26 to 35	51 to 60	16 to 25	<5

For all of the above subgroups except American Indian, the number of students in the group is more than five. Therefore, the percent proficient or above is reported. Because there are fewer than five American Indian students, the percent proficient or above is not reported. In addition, the actual number of students is not reported. In this way, it becomes far more difficult to deduce the percent or number of American Indian students scoring proficient or above. If actual numbers of students in each subgroup were reported, it might become possible, using numbers in groups and percentages, to discern confidential information. In that situation, more cells would have to be de-identified. This method allows for the maximum amount of information to be reported while still protecting the privacy of individuals.

Assessment scores can also be reported using top/bottom coding. Here, the issue is reporting information about how well a subgroup performed without revealing the exact scores of that group. If a range is reported rather than specific score levels the purpose (how the group did on the test) is met, but individual scores cannot be determined. Note that this is especially important at the top and bottom of the scale (scores of zero or 100). See Table 4.

Table 4: Top/Bottom Coding						
	Total Students	Score Range				
		>94	75-94	50-74	25-49	<25
Percent of Total	100	13	35	26	22	4
Number of Students in Subgroup	115	15	40	30	25	5

As noted earlier, if this particular subgroup were small, and the average score were 100, it would be obvious that all students earned a score of 100. If, however, a score level of >94 was reported, even if all subgroup students scored in that category, it would be impossible to determine an individual's score.

The reported score range or number of students reported in a group range would depend upon the total number of students in the group. The following could be considered for implementation of the above rules if six or more were used as the number of students in a subgroup for confidentiality purposes. See Table 5.

Table 5: Recommended Ranges for Obfuscating Actual Values		
If Total Number of Students is…	**Use Percent Above Cut-Point Intervals of…**	**Use Ranges of Number of Students of…**
<6	None	None
6-20	10	25
21-33	5	20
>33	3	5

These statements have been summarized from the review of methodologies used by statistical agencies for de-identifying the values of small groups and their relevance to education.

1. From a pure and simple statistical perspective, a minimum subgroup size of three protects the identity of the subgroup's members (degrees of freedom = 2). For example, knowing the value for one member of the subgroup still leaves two values unknown, so the value of any one of the other two cannot be determined. An example of a situation that contradicts the use of three as a minimum is a subgroup containing twins. The family of these two students would know the values for two rather than just one student.

2. Most state education agencies, school districts, and other types of agencies exceed this minimum "to be cautious." This protects against someone knowing the values of more than one student in a subgroup.

3. A minimum cell size of five will meet the requirements of confidentiality, exceed the statistical minimum of three, and provide states a comfort zone above that minimum. See Table 6.

4. Minimum cell sizes above five may inappropriately reduce the number of subgroups for which a school is responsible. Excessively high minimums will violate the intent of accountability systems by excluding subgroups and the individual students in them from accountability mandates.

Table 6: Minimum Subgroup Size of Five (5) for Confidentiality

GROUP:	All Students	White	African American	Hispanic	Asian Pacific Islander	American Indian	LEP	IEP	Economically Disadvantaged
% Proficient or Advanced	68%	20%	80%	60%	100%	100%	0%	33%	25%
Number Assessed	22	5	5	5	2	5	4	6	8
Met 75% Annual Objective?	No	No	Yes	No	Yes	Yes	No	No	No
Reported Status	Not met	Not Met	Met	Not Met	Too Few to Report	Met	Too Few to Report	Not Met	Not Met
NOTE: This table is irrespective of statistical reliability decisions.					Statistics Not Reported Publicly				

5. For reporting, if a small *n* is present, blanking out that cell in a table may not be an adequate solution. The cell value may be restorable based upon the values of other cells that are reported. See Table 7.

Table 7: Reconstituting De-Identified Cell Values									
GROUP:	All Students	White	African American	Hispanic	Asian Pacific Islander	American Indian	LEP	IEP	Economically Disadvantaged
% Proficient or Advanced	68%	20%	80%	60%	***100%***	100%	0%	33%	25%
Number Assessed	22	5	5	5	***2***	5	4	6	8
Met 75% Annual Objective?	No	No	Yes	No	***Yes***	Yes	No	No	No
Reported Status	Not met	Not Met	Met	Not Met	Too Few to Report	Met	Too Few to Report	Not Met	Not Met
NOTE: This table is irrespective of statistical reliability decisions.					Statistics Not Reported Publicly			***Values That Can be Calculated***	

6. If a school has a small subgroup, blanking out that subgroup and all others that might be used to derive that subgroup's value could result in the loss of all subgroups. This should be unacceptable in an accountability system. See Table 8.

Table 8: Loss of Valid Cells to Avoid Disclosing De-Identified Cell Values									
GROUP:	All Students	White	African American	Hispanic	Asian Pacific Islander	American Indian	LEP	IEP	Economically Disadvantaged
% Proficient or Advanced	68%	***20%***	***80%***	***60%***	***100%***	***100%***	0%	33%	25%
Number Assessed	22	***5***	***5***	***5***	***2***	***5***	4	6	8
Met 75% Annual Objective?	No	***No***	***Yes***	***No***	***Yes***	***Yes***	No	No	No
Reported Status	Not met	Not Met	Met	Not Met	Too Few to Report	Met	Too Few to Report	Not Met	Not Met
NOTE: This table is irrespective of statistical reliability decisions.					Statistics Not Reported Publicly			***Values That Can be Calculated***	
Values De-Identified to Avoid Calculation of De-identified Values									

7. As an alternative to blanking out all subgroups when one is too small to report, the values can be reported in ranges (with ranges for the n's as well) that obfuscate the actual values enough to prevent calculations. See Table 9.

Table 9: Loss of Valid Cells to Avoid Disclosing De-Identified Cell Values

GROUP:	All Students	White	African American	Hispanic	Asian Pacific Islander	America n Indian	LEP	IEP	Economically Disadvantage d
% Proficient or Advanced	68%	***0 to 20%***	***80 to 100%***	***40 to 60%***	***100%***	***80 to 100%***	0%	33%	25%
Number Assessed	22	***5 to 20***	***5 to 20***	***5 to 20***	***2***	***5 to 20***	4	6	8
Met 75% Annual Objective?	No	***No***	***Yes***	***No***	***Yes***	***Yes***	No	No	No
Reported Status	Not met	Not Met	Met	Not Met	Too Few to Report	Met	Too Few to Report	Not Met	Not Met

NOTE: This table is irrespective of statistical reliability decisions.	Statistics Not Reported Publicly	***Values That Can No Longer be Calculated***

Values De-Identified to Avoid Calculation of De-Identified Values

Conclusion

Let's return to the paradox that created the controversy. FERPA reflected back in 1974 a growing awareness that education agencies were gathering revealing data about students. Test scores are often at the center of that concern because they carry their own controversies about access, use, and disclosure. Technology has expanded the issues and processes surrounding FERPA. As a consequence of all this, the external researcher becomes an endangered species. How many education agencies have the solid data governance policy and structure in place to oversee both the protection of personally identifiable data and the need to support quality research and evaluation for program and instructional improvement and accountability?

References

- *2000 Disaggregated Achievement Report Guide Sheet.* (2000). Florida Department of Education, Curriculum, Instruction & Assessment, Evaluation & Reporting.
- Alker, H. R., Jr. (1975). Polimetrics: Its Descriptive Foundations. In F. Greenstein, and Polsby, N. (Ed.), *Handbook of Political Science. Reading: Addison-Wesley.*
- American Institutes for Research (2002). National School-Level State Assessment Score Database. (Compact Disc).
- Chou, F. (2002). Louisiana Department of Education, personal communication.
- Clements, B. S. (1998). *Protecting the Confidentiality of Education Records in State Databases*: Evaluation Software Publishing, Inc., Study for the Massachusetts Department of Education.
- Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement,* 57, 373-399.
- Evans, T., Zayatz, L., & Slanta, J. (1996). *Using Noise for Disclosure Limitation of Establishment Tabular Data:* U. S. Bureau of the Census.
- Fienberg, S. E., Steele, R. J., & Makov, U. (1997). *Statistical Notions of Data Disclosure Avoidance and Their Relationship to Traditional Statistical Methodology: Data Swapping and Loglinear Models:* Carnegie Mellon University, Haifa University.
- Gravetter, F. J., & Wallnau, L. B. (2000). *Statistics for the Behavioral Sciences* (Fifth ed.). Belmont, CA: Wadsworth/Thomson Learning.
- Gurr, T. R. (1972). *Politimetrics: An Introduction to Quantitative Macropolitics.* Englewood Cliffs, NJ: Prentice-Hall.
- Hambleton, R. K., & Slater, S. C. (1997). Reliability of Credentialing Examinations and the Impact of Scoring Models and Standard-Setting Policies. *Applied Measurement in Education*, 10(1), 19-38.
- Hays, W. L. (1994). *Statistics*. Austin: Harcourt Brace College Publishers.
- Hilton, G. (1976). *Intermediate Politometrics.* New York: Columbia University Press.
- Hoffman, R. G., & Wise, L. L. (2000). *School Classification Accuracy Final Analysis Plan for the Commonwealth Accountability and Testing System.* Alexandria, VA: Human Resources Research Organization (HumRRO).
- Jaeger, R. M., & Tucker, C. G. (1998). *Analyzing, Disaggregating, Reporting, and Interpreting Students' Achievement Test Results: A Guide to Practice for Title I and Beyond.* Washington, D.C.: Council of Chief State School Officers.
- Kane, M. (1996). The Precision of Measurements. *Applied Measurement in Education,* 9(4), 355-379.
- Kim, J. J., & Winkler, W. E. (1997). *Masking Microdata Files*. Paper presented at the American Statistical Association.
- King, G. (1991). *On Political Methodology.* Paper presented at the American Political Science Association, Atlanta, Georgia.
- Ligon, G. D., Clements, B. S. (2008). Revisions to FERPA Guidance. ESP Solutions Group, Inc.
- Ligon, G. D. (1998). *Small Cells and Their Cons (Confidentiality Issues):* NCES Summer Data Conference.
- Ligon, G. D., Clements, B. S., & Paredes, V. (2000). *Why a Small n is Surrounded by Confidentiality: Ensuring Confidentiality and Reliability in Microdatabases and Summary Tables*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.
- Ligon, G.D., Jennings, Judy, & Clements, B.S. (2002). *Confidentiality, Reliability, and Calculation Alternatives for No Child Left Behind.* Unpublished paper for CCSSO CAS/SCASS.
- Linn, R. L., & Haug, C. (2002). Stability of School-Building Accountability Scores and Gains. *Educational Evaluation and Policy Analysis*, 24(1), 29-36.
- Moore, R. A. (1997). *Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets*. Unpublished manuscript, Washington, D C.

- Newton, Isaac. (1687). Mathematical Principles of Natural Philosophy *North Carolina Student Accountability Standards*. North Carolina Department of Public Instruction. Retrieved, from the World Wide Web: www.dpi.state.nc.us/student_promotion/sas_guide/standard_error.html
- Opperdoes, F. (1997). *Bootstrapping*. Retrieved, from the World Wide Web: http://www.icp.ucl.ac.be/~opperd/private/bootstrap.html
- Rai, K. B., & Blydenburth, J. C. (1973). *Political Science Statistics*. Boston: Holbrook Press.
- Rogosa, D. (1999). *Statistical Properties of Proportion at or above Cut-off (PAC) Constructed from Instruments with Continuous Scoring.* UCLA: National Center for Research on Evaluation, Standards, and Student Testing.
- *Statistical Policy Working Paper 2--Report on Statistical Disclosure and Disclosure-Avoidance Techniques*. (1978). Federal Committee on Statistical Methodology.
- Winkler, W. E. (1997). *Views on the Production and Use of Confidential Microdata:* Bureau of the Census.
- Yen, W. M. (1997). The Technical Quality of Performance Assessments: Standard Errors of Percents of Pupils Reaching Standards. *Educational Measurement*: Issues and Practices (Fall), 5-15.
- Zayatz, L., Moore, R., & Evans, B. T. (undated). *New Directions in Disclosure Limitation at the Census Bureau*. Washington, DC: Bureau of the Census.

EXHIBIT A

HIPAA Methods for De-Identification

Implementation specifications: requirements for de-identification of protected health information: A covered entity may determine that health information is not individually identifiable health information only if:

1. Expert Determination

A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable: (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and (ii) Documents the methods and results of the analysis that justify such determination.

or

2. Safe Harbor

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names	
(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000	
(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older	
(D) Telephone numbers	(L) Vehicle identifiers and serial numbers, including license plate numbers
(E) Fax numbers	(M) Device identifiers and serial numbers
(F) Email addresses	(N) Web Universal Resource Locators (URLs)
(G) Social security numbers	(O) Internet Protocol (IP) addresses
(H) Medical record numbers	(P) Biometric identifiers, including finger and voice prints
(I) Health plan beneficiary numbers	(Q) Full-face photographs and any comparable images
(J) Account numbers	(R) Any other unique identifying number, characteristic, or code, except as permitted by the section “Re-identification”; and
(K) Certificate/license numbers	

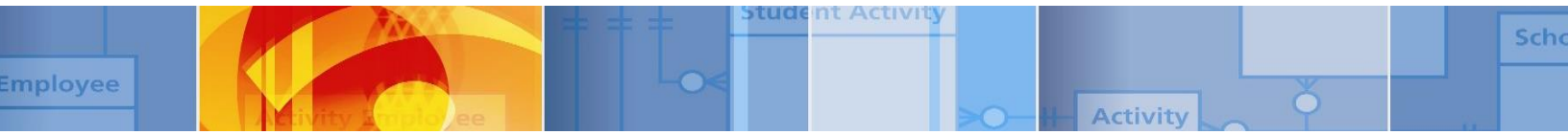
(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

Satisfying either method would demonstrate that a covered entity has met the standard in §164.514(a) above. De-identified health information created following these methods is no longer protected by the Privacy Rule because it does not fall within the definition of PHI.

Re-identification

A covered entity may assign a code or other means of record identification to allow information de-identified under this section to be re-identified by the covered entity, provided that:

(1) *Derivation.* The code or other means of record identification is not derived from or related to information about the individual and is not otherwise capable of being translated so as to identify the individual; and
(2) *Security.* The covered entity does not use or disclose the code or other means of record identification for any other purpose, and does not disclose the mechanism for re-identification.



About ESP Solutions Group

ESP Solutions Group provides its clients with *Extraordinary Insight*™ into P20W education data systems and psychometrics. Our team is comprised of industry experts who pioneered the concept of "data-driven decision making" and now help optimize the management of our clients' state and local education agencies' information systems.

ESP personnel have advised school districts, all state education agencies, and the U.S. Department of Education on the practice of P20W data management. We are regarded as leading experts in understanding the data and technology implications of NCLB, Access4Learning/SIF, ED*Facts*, CEDS, state reporting, metadata standards, data governance, data visualizations, and emerging Edtech issues**.**

Dozens of education agencies have hired ESP to design and build their longitudinal data systems, state and federal reporting systems, metadata dictionaries, evaluation/assessment programs, and data management/analysis and visualization systems.

To learn how ESP can give your agency *Extraordinary Insight* into your P20W education data, contact us at (512) 879-5300 or info@espsg.com.

This document is part of *The Optimal Reference Guide* Series, designed to help decision makers analyze, manage, and share data in the 21st Century

(512) 879-5300
www.espsolutionsgroup.com