



The ESP Journal

Extraordinary insight into today's education information topics

Privacy versus Access to Data for School Improvement

Glynn D. Ligon, Ph.D.

Updated December 2021

Based upon a Paper Presented at the Annual Meeting
of the American Educational Research Association



Table of Contents

Abstract 1

Setting the Stage 2

EXHIBIT A HIPAA Methods for De-Identification..... 13

EXHIBIT B Definition of the Sets..... 15

EXHIBIT C POSTER SUPPLEMENT..... 32

References 42

Abstract

Education agencies must follow FERPA and HIPAA to provide parents access to student data while protecting unauthorized public access. However, agencies too often over-react in their stewardship and limit data for analytics that could result in school improvement or appropriate public awareness.

Researchers and evaluators can be frustrated to find that crucial data elements are missing from student records provided by education agencies. De-identification to protect personally identifiable information removes key analytical characteristics of individuals. This study identified how education agencies manage their own databases and those for the public and researchers. Recommendations are made for a specific set of data to be created with identified records for authorized and authenticated researchers and evaluators.

A four-tier structure is proposed to provide internal, historical data; limited access for analytics; and fully public statistics.

Setting the Stage

FERPA is often misunderstood and misapplied by education agencies and even their legal advisors. FERPA doesn't deny researchers and evaluators access to students' personally identifiable information (PII). FERPA doesn't mandate that education agencies provide external researchers and evaluators only de-identified databases. However, to protect personally identifiable information, some education agencies may have de-identified or masked so many data elements within their databases that research and evaluation (RE) is at risk. These education agencies have defaulted to this strategy as the ultimate protection against unintentionally revealing confidential data. They may also have adopted this policy as the easiest or most equitable response to the quantity of requests they receive for data. Unfortunately, this stance may also be the result of a lack of a full understanding of FERPA and the mutually beneficial alternatives available for creating sets of data that can meet everyone's needs.

Set: A dataset or database containing a group, or subset, of data elements. The elements may be granular within individual records or aggregate statistics in entity records or within tables. The specific elements in a set are selected to match the needs and authorized uses of the persons who will be provided access to the dataset.

FERPA (Family Educational Rights and Privacy Act): The 1974 federal legislation that is the foundation for allowing parents access to and protecting student information in education records

PII (Personally Identifiable Information): Information or data elements that can be used on their own or with others to identify, contact, or locate a single person, or to identify an individual in context

RE (Research and Evaluation): Within the context of this paper, investigative studies testing hypotheses about education, instruction, program development, assessment, and other issues broadly informed using data available from schools and education agencies

This paranoia by education agencies and the over-de-identification process it creates either prevents the collection of necessary data elements or removes them from the student records available to researchers and evaluators. At times, the de-identification is even ineffective. Ineffective in the sense that a clever analyst can recover the data from marginal totals and other remnants left in the database or related tables. Other times, it is unnecessary. Unnecessary in the sense that the elements removed were not actually personally identifiable. Almost always, as this paper describes, there is a better

methodology available.

This study is applied research into the problem of how agencies can protect PII while preserving the integrity of data for RE. In this case, the problem studied is both institutional for education agencies and professional best practice for researchers and evaluators. This places the issue squarely in the charter of AERA's Division H. The audience for privacy, confidentiality, de-identification, or RE using school district, state education agency, and university-based databases has grown to include almost everyone in the educational research community. Meanwhile, the RE profession has been championing the protection of PII often to its own disadvantage.

Applied Research: Study that seeks to solve practical problems

Significance of This Problem: An education agency has two challenges when de-identifying the records within a database. Simultaneously, the agency should:

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

As soon as an agency sets out to establish the business rules for de-identifying individual records, the paradox of these goals becomes apparent. Removing the PII of individuals degrades the integrity of the records for analysis and reporting. For example, a researcher requests a database to study gender and race differences for military connected students. The Every Student Succeeds Act (ESSA) requires states to identify and report students whose parents have military connections. The education agency may read FERPA and remove the students' names, addresses, and birthdates because those are PII. Then they remove the students' race, ethnicity, gender, and enrollment dates because those are conditionally PII data elements. The researcher's resultant analyses will not be very precise.

Conditional PII: A data element that on its own is not personally identifiable but when combined with others becomes personally identifiable

Isaac Newton's Third Law of Motion: For every action, there is an equal and opposite reaction (Mathematical Principles of Natural Philosophy, 1687) (46)

Let's restate Isaac Newton's Third Law of Motion 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.

Take for example the case of John Snow back in 1854 who ended a cholera outbreak by

noticing that the addresses of victims centered around a water pump. (47) What if HIPAA and FERPA had prevented the provider of his database from revealing the personally identifying data element of address for the individuals in Chicago? Imagine if a future example is the discovery that students born within a specific date range in a certain city perform significantly better in a phonics program; however, that finding is never made because birthdate and address were masked within the research database provided the program evaluators? Therefore, students are never regrouped to their academic advantage. How do we resolve the paradox of faithfully performing our data stewardship to protect PII while preserving the data integrity within databases for RE? If they cannot assure politicians and parents that PII is secure and confidential, then education agencies are at risk of losing the authority to collect and store those data as in Louisiana. (1) In that state, a restrictive law was passed reducing the authority of the state education agency to collect PII. If that occurs, then researchers and evaluators, both internal and external, lose the ability to conduct the studies that contribute to school improvement. The schools may also have gaps in their accountability reporting of official statistics to the public, the state, and the federal government.

The Senate's 2015 Student Privacy Protection Act proposed to prohibit funding for any agency that "appends" PII through data matches, and any student data from being included in state longitudinal data sets unless they are first "aggregated, anonymized, and de-identified." AERA Executive Director Felice J. Levine protested, "At the heart of our concern is to avert putting student privacy and the quality of student data on a collision course." Ultimately, the Every Student Succeeds Act (ESSA) said little about PII, deferring to a future update of FERPA (2, 3, and 4).

Division H is AERA's champion for conducting the applied research to inform issues like this one. For local education agencies (LEAs), the issue is very basic: Will parents opt out of providing data (e.g., participation in assessments, surveys, studies, directory information, etc.)? How can we assure them our systems are confidential and our governance processes protect their children's PII? When responding to requests from external researchers, does a de-identified database support RE findings that improve instruction and support decision making by the LEA? Are new state privacy laws reasonable or reasonably implemented? (5 and 6) What practice is successful now on each side of the issue? Is the best solution a compromise, i.e., a compromised database? Does this sacrifice some degree of confidentiality along with some sacrifice of desired data elements? Currently, education agencies are too often choosing either a compromise that underserves researchers and evaluators, or choosing to avoid all risk by de-identifying and not accommodating RE at all.

Theoretical Framework: RE must work with policy makers to achieve mutual goals. They both must work with information technology (IT) professionals as well. This study crafted a theoretical framework of politometrics and polititech to design a solution model that

researchers and evaluators, policy makers, and IT professionals can endorse.

IT: IT, Information Technology, the development, maintenance, and use of computer systems, software, and networks for the processing and distribution of data/information

Politimetrics: Decisions made through science (psychometrics) and policy (politics)

Examples are adopting a proficiency score, how many credits to require for graduation, and what cell size protects confidentiality. Neither the pure psychometricians nor the pure politicians should make these decisions independent of the data and wisdom of the other. (7)

Polititech: Merging of policy (politics) and technology to create data governance policies and processes

Polititech is important because data governance is the solution to resolving the paradox of protecting PII while maintaining the integrity of the contents of the longitudinal data system for analytics. (8)

Data governance, which includes managing PII into and out of databases, is quintessential polititech. Designing a database model to fit the political (e.g., FERPA and HIPAA) mandates, the compliance reporting rules of enabling legislation, and the analytical requirements of researchers is polititech.

Likewise, there should be complete agreement that a single data governance policy overseeing everything is essential. Data governance (including politimetrics) is an essential process for education agencies to manage their information resources, including how RE is conducted. (9)

Data Governance: An education agency's policies and processes for overseeing the collection, storing, accessing, and reporting of its data

FERPA and HIPAA do not restrict an education 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.

Conceptual Framework: The conceptual framework for this study was to (A) describe extant models of education agencies publishing data and responding to data requests, (B) evaluate them for efficacy related to both protection of PII and support of RE, and (C) identify/develop an architecture that satisfies both requirements. This framework was supported by a review of longitudinal data systems, their associated data governance

policies, and literature on de-identification of PII. The de-identification methodologies implemented within the data systems and the masking techniques employed during the reporting processes were documented. This provided the content for the

conceptual framework for understanding and designing strategies that successfully addressed the stated problem.

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.

Study Design: The study design incorporated a review of the literature of confidentiality, de-identification, and data model architecture. Viable solutions for education agencies that ensure confidentiality for PII were identified. The Office of Management and Budget issued their “Statistical Policy Working Paper 22 - Report on Statistical Disclosure Limitation Methodology. Federal Committee on Statistical Methodology” (13), which describes in detail multiple methods. The National Center for Education Statistics (NCES) has issued several guidance documents with methods for education agencies. (10) (11) (12) (14) (15) (16)

The study team was in a unique position to access descriptions of longitudinal data systems across the nation from its direct work with nine contracts for vertical reporting (state reporting from schools/LEAs to state education agencies (SEAs) in Alaska (17), Arizona (18), Connecticut (19), Iowa (20), Idaho (21), Missouri (22), Ohio (23), Utah (24), and Wyoming (25)) and 11 other contracts to design and/or build longitudinal data systems (Alaska (26), Colorado (27), Delaware (28), Idaho (29), Louisiana (30),

Missouri (31), Montana (32), South Dakota (33), Tennessee (34), Texas--ESC Region 10 (35), and Wyoming (36)). In addition, descriptions were available of other systems funded by IES/NCES (37).

Vertical Reporting: The process of reporting data from one level of government to the next higher level (e.g., from districts to states or from states to the federal government)

Data Collection and Analysis Procedures: The study methodology involved direct examination of the metadata dictionaries of each organization to determine the PII collected and maintained in their operational data stores. Business rules for de-identification and masking for reporting were examined. This included details of the data elements reported by the LEAs and stored in the SEA databases. These data elements and the derived official statistics published in public reports were also examined. Analysis procedures used content analytics of the metadata and descriptive documentation of these systems. For this study,

the focus was mainly on the de-identification processes implemented both within the databases and upon reporting. The study investigated methodologies to discover unique processes across the agencies. The practices documented were compared and contrasted with theory and guidance provided in the literature. (38) (39) The conceptual framework was followed to produce the four-set architecture that was recommended.

Findings: The findings are presented within the theoretical framework described for RE, policy makers, and IT. IT must be able to deliver a solution that both policy makers and RE can use effectively. The first finding from the review of the literature, relates to the analysis of HIPAA's guidance and its relationship to PII in education records.

HIPAA: The Health Insurance Portability and Accountability Act (1996) that protects health records

HIPAA guidance (included in Exhibit A) simplistically specifies 18 data elements to remove from a database to achieve safe harbor status. (40, 41) With those data elements gone, a database is considered to protect identities. Unfortunately, suppressing all these data elements did not suffice for these education agencies. The concept that introduces the most question is that of conditionally identifiable elements.

Conditional PII: Data elements that on their own are not PII, but become personally identifiable when known in combination with other data elements

FERPA defines these as elements that when combined with other elements can identify the identity of an individual. Gender, race/ethnicity, and disability conditions are not first-line PII in FERPA; however, they may be defined, and are frequently by education agencies, as conditional PII elements. These are not among HIPAA's safe harbor elements. They are elements education agencies deem linkable to individuals.

Many of the 18 data elements are typically essential co-variables or classification variables in education studies. Suppressing them in a database is a disabling methodology from the perspective of RE.

The review of governance practices for extant systems built by SEAs from the data collected from LEAs showed these practices for protecting PII when responding to requests for data.

1. Refusal of Requests

- Not all agencies had clear data governance policies describing the application, review, approval, and appeal process. Controlling and protecting PII was managed at times simply by not approving requests for data.

2. Redacted Reports

- For one-time, ad hoc requests, an efficient response is to provide an

- unpublished report with confidential sections redacted.
3. Tables with Small Cells Masked
 - A common practice is to follow standard masking techniques for small cells. This is the practice evident on agencies' public websites.
 4. Ad Hoc Responses
 - Requests that require custom analyses generate unique responses—and are considered “one-time” activities that are not posted for other audiences to share. These do not become standard processes or reports.
 5. Ad Hoc De-Identified Databases
 - Requests for records that contain PII require custom de-identification. These are typically handled through the agency's research request review and approval process.
 6. De-Identified Research Databases
 - Some agencies have a prepared de-identified database available for researchers. Even these may be accessible within a very controlled environment. For example, users may be required to access them within a controlled lab environment.
 7. Extract of Identified Records
 - Authorized researchers with approved research studies infrequently are provided a database.
 - Technically, some precoding databases provided to assessment vendors are in this category.
 8. Authorized, Authenticated Access to Identified Internal Databases
 - This was only found to be infrequently available to limited vendors and individuals under contract with the agency after signing a confidentiality statement.

These eight responses should be guided by the agency's data governance policy. There are three fundamental processes within an education agency that the data governance policy must support.

1. Operations and Official Reporting from the Agency's Authoritative Data Source(s) (e.g., human resources, finance, student information system, longitudinal data system, etc.)

The agency must have operational data systems with their authoritative sources of data. These data would support on-going operations with real-time and longitudinal data— unmodified records for official purposes and reporting. For security, these systems would be behind a firewall and inaccessible to unauthorized personnel. Researchers and evaluators would not access these raw data sources.

Authoritative Data Source: The single set of data upon which the agency depends to be its official record and to which all other data are compared for accuracy

Operational Data System: The data system that runs to day-to-day operations and functions of the agency, e.g., payroll, grade reporting, bus routing, etc.

Longitudinal Data System (LDS): Data warehouse, database, datastore, datamart, 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, the P20W SLDS designates a multiagency statewide system. W refers to the workforce.)

Research and Evaluation

The agency must accommodate four classes of researchers and evaluators.

The first class includes their own internal employees with a need to know based upon their position and role within the organization.

The second class is external persons who request permission, in compliance with the agency's Governance Policy, FERPA, and all other applicable regulations, to conduct a study requiring access to data that includes PII.

The third class is external persons who request permission, in compliance with the agency's Governance Policy, FERPA, and all other applicable regulations, to conduct a study requiring access to data that does not include PII.

- a. The fourth class is persons making requests under freedom of information provisions. These would receive records without PII.

To be efficient and compliant, the agency must have a de-identified database to provide with confidence to external researchers in the third and fourth classes. 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.

2. Public Reporting

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.

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 where it is avoidable.

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. Authenticate the users upon accessing the data.
3. Authorize the users for approved purposes.
4. Mask the data in any small, reported fields or cells.

Whatever the choice, the agency's Data Governance Policy should protect the PII within an 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 study's guidance is to de-identify a database for RE purposes when 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.

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, 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.

Overall Contribution to the Field: This study provides education agencies an architecture to respond to both FERPA's mandate to protect PII and RE's requirement for internal integrity in databases. Education agencies that now over-suppress the data within their longitudinal databases and as a consequence disable the ability of REs to conduct useful studies can follow this blueprint to implement a four-set architecture. The contributions to the field from this study are twofold. The data governance processes in an education agency that must be in place to satisfy the requirements of FERPA/HIPAA are demonstrated. The characteristics of confidential databases viable for RE have also been defined.

The overall contribution of this study is a description of a four-set architecture that is responsive to the requirements for protecting PII and maintaining analytic integrity. Exhibit B provides descriptions for each of these sets. Each is described by the users targeted, some sample questions they might ask, and the typical de-identification methods that could be employed.

Set 1: Authoritative Data Source(s)—Core data system containing identified records for operations and official statistics and reports

Internal RE is conducted using these fully identified data. Agency staff use these data for official statistics and accountability.

Set 2: Research and Evaluation Database—Individual records selected for longitudinal analysis with identities and demographics provided

Authorized and authenticated external researchers may access these data (#7 above). Security measures assure that PII are safe and confidential.
(42)

Set 3: De-Identified Research Database—Available to researchers and persons making open- records requests

Data governance policy guides how users access these data through portals or downloads (#5 and 6 above). Record Code Substitution (Tokenization) (43) allows re- identification where appropriate for matching of records across years to longitudinal records and across files for analyses across programs and agencies. Data suppression (removing data) degrades the research integrity of the database. More sophisticated statistical disclosure control methods perturb the data (e.g., MASSC (44) and multiple imputation (45)).

Set 4: Publicly Reported Statistics with Small Cells Masked (Aggregate Public Reporting Data Portal)—Reports with aggregate statistics, small cells masked

Data portals may be implemented in many ways to present reports (#3 and 4 above). Masking (reconfiguring data) techniques include categorization of continuous variables, substituting individual values for group averages, controlled rounding, combining cells, top/bottom coding, redaction, disclosure avoidance (denying data), and disapproval of requests. Business rules are enforced to ensure cells with too few individuals are not reported and cannot be recalculated from other cells.

Recommendations for future research include:

- Comparing security and confidentiality rules of Set 2 databases;
- Comparing the efficacy of Set 3 de-identification rules used by different agencies, especially when additional years of data are added to longitudinal databases; and
- Comparing the effectiveness of masking strategies for Set 4 reporting methods.

Summary and Conclusions

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 RE for program and instructional improvement and accountability?

Therefore, Newton's Third Law of Motion for databases was expanded to be that every action taken to de-identify a database creates an equal and opposite reaction against our capability to analyze and report from that database. Reactions within databases occur exponentially. Thus, deleting a single data element can in reality disable untold combinations and permutations of relationships and causalities discoverable by REs.

Researchers and evaluators must become engaged in the data governance of education agencies to ensure that the research integrity of databases is protected. The public, parents, and politicians should be satisfied with security controls protecting PII inside a Set 1 and Set 2 database, which allow quality data with the integrity required for operations, official statistics, and valid RE designs. A Set 3 de-identified research database should satisfy confidentiality concerns while providing data for many analytic and informational purposes. A Set 4 aggregate public reporting data portal should provide public reporting of official statistics with protection of PII. This four-set solution of enhanced statistical and technical methodologies can be successfully implemented within a data governance system that employs the wisest politometrics and polititech to protect both PII and the integrity of data for RE.

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.

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

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: The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and
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.

EXHIBIT B Definition of the Sets

Set 1: Authoritative Data Sources(s)

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 store. 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).

Authoritative Data Source(s)		
Users	Internal	Program Officers, IT Staff, Agency Officials
	External	Authenticated & Authorized Researchers, Contractors

Questions	Internal	<ul style="list-style-type: none"> • What are our agency's official statistics? • What students meet early warning criteria? • What schools met annual accountability objectives? • How many busses are needed each day? • How many lunches were served; full price, reduced, free? • What is the fund balance for the fiscal year?
	External	<ul style="list-style-type: none"> • 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

Set 2: Research and Evaluation Database

Research and Evaluation Database		
Users	Internal	RE Staff
	External	Authenticated & Authorized Researchers
Questions	Internal	<ul style="list-style-type: none"> • What modifications to the current growth model would improve the accountability system? • Which LMSs have the best ROI?
	External	<ul style="list-style-type: none"> • Are my dissertation hypotheses supported? • Are English language learners migrating into or out of the inner city?
De-Identification Methodology		None

Set 3: De-Identified Research Database

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 Research Database		
Users	Internal	RE Staff
	External	Researchers, FOI Requestors
Questions	Internal	<ul style="list-style-type: none"> Have statewide performance level trends changed?
	External	<ul style="list-style-type: none"> Did X Reading Program outperform Y Reading Program statewide? What was the impact of changes in graduation requirement policies statewide?
		<ul style="list-style-type: none"> Has enrollment in charter schools changed?
De-Identification Methodology		<ul style="list-style-type: none"> Safe Harbor Expert Determination <ul style="list-style-type: none"> 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.

Set 4: Publicly Reported Statistics with Small Cells Masked (Aggregate Public Reporting Data Portal)

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.

Publicly Reported Statistics with Small Cells Masked (Aggregate Public Reporting Data Portal)		
Users	Internal	All Staff
	External	Researchers, All External Data Users
Questions	Internal	<ul style="list-style-type: none">• Have statewide performance level trends changed?
	External	<ul style="list-style-type: none">• 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		<ul style="list-style-type: none">• 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 development 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 (Table 1), and (3) top/bottom-coding (Table 2).

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.”

Table 1: Limitation of Detail Using Categories/Ranges for Number of Students						
	Total Students	African American	Hispanic	White	Asian	American Indian
Percent 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

Table 2: 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

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	American Indian	LEP	IEP	Economically Disadvantaged
% 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									

EXHIBIT C POSTER SUPPLEMENT

Privacy versus the Integrity of Research and Evaluation in Schools

AERA Annual Meeting, San Antonio, April 27, 2017

Poster Session 3 Applied Research in Schools: Education Policy and School Context
Glynn D. Ligon, Ph.D. Evaluation Software Publishing, Inc., Austin, Texas
gligon@espsg.com

Disclaimer: The characters and characterizations used in this supplement are intended to be illustrative—not literal. FERPA does not actually apply to the identities of comic book characters, does it?

“Who was that masked man?”

That question was asked over 3,000 times at the end of episodes of a familiar radio, TV, and movie series. The rancher whose property had just been saved from outlaws would turn to his neighbor and say, “**Oh, he’s the Lone Ranger.**”

However, if John Reid (the Lone Ranger) were a public school student and FERPA had been in force, the rancher would have added, “**...and his real name and face are masked because they are personally identifiable information.**” Simply put, the Lone Ranger’s mask hid his identity back in the 1930’s similar to how education agencies use identification numbers and de-identification techniques following FERPA mandates beginning in the 1970’s.



Another Example If Jean Grey, the metamutant X-Men super hero, saved your child on the playground, her school wouldn’t tell you her real name was Jean Elaine Grey, who first appeared in X-Men #1 in 1963, and she’s ranked #13 on IGN Entertainment’s list of super heroes. Her parents, Elaine and John, raised her in Annandale-on-Hudson, New York until she began protecting Alphabet City. In fact, they wouldn’t even confirm that she’s female and white—or that she is the figment of the imagination of Stan Lee and Jack Kirby. That’s because any of those data elements alone or in combination with others might allow you to identify Jean and then discover other information about her in their database or published statistics.

Without personally identifiable data elements, a researcher interested in super heroes would be unable to answer fascinating questions such as...

Does the Body Mass Index (BMI) of super heroes differ significantly by gender, race, and ethnicity?

- I hypothesize that it does. My preliminary analysis is confirmatory.
 - Super heroes may reflect their creators and illustrators’ stereotypes.

Super Hero Examples Continuing with the super hero theme, I researched 10 individuals. Figure 1 presents their actual records from this database. Typical of an education database, there are some missing data that were never revealed in the super heroes' publications. To ensure as complete and accurate data as possible for each one, I searched sources on the Internet, then visited Dragon's Lair (2438 West Anderson Lane, Suite B1, Austin, TX 78757, 512-454-2399). Their resident expert, Bobby, was generous in verifying the existing data and agreeing that certain data are missing from recorded history. He was able to fill in a few fields I could not.

Figure 1: Individual Records

PI?	Data Element	Super Hero									
P	Super Hero Name	Batman	Black Lightning	Catwoman	Jean Grey	Katana	Lone Ranger	Rafael	Robin	Spider-Man	Tonto
P	Real Name	Bruce Wayne	Jefferson Pierce	Selina Kyle	Jean Elaine Grey	Tatsu Yamashiro	John Reid	Rafael	Dick Grayson	Miles Morales	Tonto
P	Date of 1 st Appearance	1939	1977	1940	1963	1983	1933	1984	1940	2011	1933
P	1 st Publication	Detective Comics #27	Black Lightning #1	Batman #1	The X-Men #1	Brave & the Bold #200	Detroit Radio WXYZ	Nina Turtles #1	Detective Comics #38	Ultimate Fallout #4	Detroit Radio WXYZ
P	Mother's Name	Martha Wayne	Mrs. Pierce	Maria Kyle	Elaine Grey	Missing	Missing	Missing	Mary Grayson	Rio Morales	Missing
P	Father's Name	Thomas Wayne	Alvin Pierce	Brian Kyle	John Grey	Missing	Missing	Missing	John Grayson	Jefferson Davis	Potawatomi Chief
P	IGN Super Hero Rank	2	85	20	13	Unranked	Unranked	23	11	3	Unranked
C	Type	Human	Human	Human	Metamutant	Human	Human	Mutant Turtle	Human	Human	Human
C	Birthplace	Gotham City	Suicide Slum, Metropolis	Gotham City	Annandale-on-Hudson	Missing	Missing	New York	Gotham City	Brooklyn, New York	Potawatomi Nation
C	Home City	Gotham City	Suicide Slum, Metropolis	Gotham City	Annandale-on-Hudson	Missing	Missing	New York	Gotham City	New York	Missing
C	Protected City	Gotham City	Metropolis	Gotham City	Alphabet City	Gotham City	Missing	New York	Gotham City	New York	Missing
C	Gender	Male	Male	Female	Female	Female	Male	Male	Male	Male	Male
C	Race	White	African American	White	White	Japanese	White	Reptile	White	African American, Puerto Rican	Native American, Comanche
C	Hispanic	No	No	No	No	No	No	No	No	Hispanic	No
C	Creator Name(s)	Bob Kane, Bill Finger	Tony Isabella, Trevor Von Eeden	Bob Kane, Bill Finger	Stan Lee, Jack Kirby	Mike W. Barr, Jim Aparo	Fran Striker	Kevin Eastman, Peter Laird	Bob Kane, Bill Finger, Jerry Robinson	Brian Michael Bendis, Sara Pichelli	George W. Trendle, Fran Striker
N	Super Powers	Weaponry	Electro-kinesis, Flight	Burglary, Gymnastics	Telepathy, Telekinesis	Martial Arts, Soul Capture	Riding, Shooting	Sai Stick	Gymnast, Martial Arts	Strength, Agility	Riding, Shooting
N	Sidekick?	Yes	No	No	No	No	Yes	Yes	No	No	No
N	Is Sidekick?	No	No	No	No	No	No	Yes	Yes	No	Yes
N	Height	6' 2"	6' 1"	5' 7"	5' 8"	5' 2"	6' 0"	5' 2"	5' 10"	5' 10"	6'
N	Weight	210	200	125	156	96	175	180	175	167	170
N	BMI, Level	27 Overweight	27 Overweight	20 Normal	24 Normal	18 Normal	23.7 Normal	32.9 Obese	25.1 Overweight	24 Normal	23.1 Normal

Abbreviations in Figure 1



PI?: Is the data element personally identifiable? The codes are:

P: Personally Identifiable—a data element that meets FERPA’s definition of identifying an individual

C: Conditionally Personally Identifiable—a data element that when combined with one or more other data elements becomes personally identifiable

N: Not Personally Identifiable—a data element that does not identify an individual

Remember That BMI Hypothesis? Looking at our 10 super heroes' BMI data in Figure 1 and graphed in Figure 2, we get a glimpse at whether there might be bias in their height and weight by gender, race, and ethnicity. But wait, the point of this paper is that we may not get to see this detail.



Katana

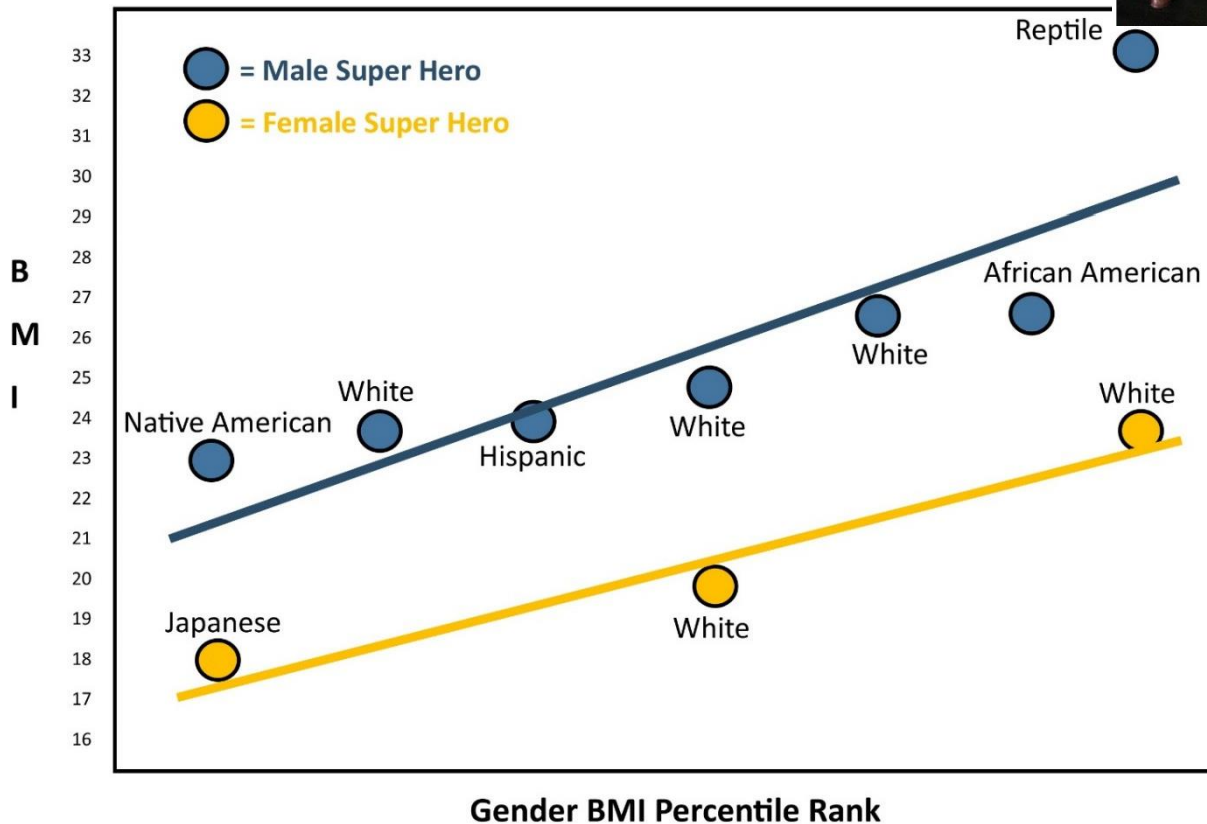


Figure 2: BMI by Gender Percentile Rank

Sets to Meet Varied Requirements The full paper recommends that education agencies create four “sets” (or databases) of data to ensure that at least one of them meets the needs of researchers and evaluators.

Set 1: Identified Internal Longitudinal Data Store(s)—Core data system containing identified records for official purposes

Internal RE is conducted using these fully identified data. Agency staff use these data for official statistics and accountability. These data are used to run the business processes of the organization.



Set 2: Identified Longitudinal Data Store for Research and Evaluation—Selected sample of identified records for unrestricted RE

Authorized and authenticated users follow established data governance guidelines to conduct approved analyses

Set 3: De-Identified Research Database—Database of records without PII and conditional information available to persons making open-records requests

Data governance policy guides how users access these data through portals or downloads. Record Code Substitution (Tokenization) allows re-identification where appropriate for matching of records across years to longitudinal records and across files for analyses across programs and agencies. Data suppression (removing data) degrades the research integrity of the database. More sophisticated statistical disclosure control methods perturb the data (e.g., MASSC and multiple imputation).

Set 4: Aggregate Public Reporting Data Portal—Reports with aggregate statistics, small cells masked

Data portals may be implemented in many ways to present reports). Masking (reconfiguring data) techniques include categorization of continuous variables, substituting individual values for group averages, controlled rounding, combining cells, top/bottom coding, redaction, disclosure avoidance (denying data), and disapproval of requests. Business rules are enforced to ensure cells with too few individuals are not reported and cannot be recalculated from other cells.



Our Heroes’ Data Let’s look at our group of 10 super heroes. They are diverse by gender, race/ethnicity, and other PII. As interesting as that would make our

social interaction with them, it makes our reporting of their data problematic. All of their disaggregated cells will be fewer than 10. Thus, if we self-impose a rule of masking small cells below 10, that will prevent our reporting anything except the grand total of 10 in our tables.

In Figure 1, the first seven data elements are P, personally identifiable—Date of 1st Appearance, for example. The next eight are conditional. These on their own may not identify the hero, but in combination with other data elements single out one of them.

Our Heroes' Data in the Four Sets Walking through each of the four sets is illustrative. Figure 1 is equivalent to Set 1—all the data elements with no restrictions. For an authorized and authenticated researcher or evaluator, Set 2 could be a duplicate of Set 1. Just for the sake of example, let's say that the organization chooses not to include in Set 2 the data elements Mother's Name and Father's Name. Those are useful to the organization, but their de-identification or deletion most likely devalues the analytic nature of Set 2 very little.



Set 3, however, would contain only the six data elements designated as not personally identifiable or conditionally personally identifiable. This would disable any analyses for Set 3 users for questions about gender, race, ethnicity, cities, dates of first appearance, rankings, or creators.

Set 4 is the interesting case. Set 4 is not a subset of the records in Figure 1, but the aggregate statistics derived from them. Set 4 official statistics for an organization should be created from Set 1. The Data Governance Policy should determine how the published statistics are masked using a small number rule. Therefore, any and all statistics can be calculated, then the small cell size rule can be imposed prior to publication. This allows more cells to be calculated and be eligible for publication than if the de-identified Set 3 is used for calculating statistics. This is a crucial distinction. This means that researchers using Set 3 typically cannot replicate all of the official statistics in Set 4.



Using our small group of 10 super heroes as an example, their PII is de-identified for users of Set 3. That doesn't just impose masking of small cells for reporting, it prevents even the initial calculating of statistics for subgroups using the data elements that have been de-identified.

Imagine now that there are 10 groups of super heroes just like ours. If they are combined into a megateam of 100, then they could be reported in subgroups large enough to meet the rule of 10. A researcher using Set 2 would be able to do this. One restricted to Set 3 would yet again have only

a grand total—now for 100. Examples of a Set 4 official statistic, Body Mass Index, are in Figures 3A, 3B, and 3C. Figure 3A shows the BMI level frequencies calculated and displayed from Set 2, our 10 heroes with no masking. Figure 3B shows the reporting for our megateam of 100 heroes with their data fully identified in Set 2. Finally, Figure 3C shows that using masked data from Set 3, only the grand total of 100 super heroes could be calculated with no breakouts.

Researchers Need PII in Set 2! This is why authorized and authenticated researchers need access to a Set 2 database. This is why a Set 3 database with both PII and conditional PII deleted probably destroys the integrity of the data for most RE questions of significance. Unfortunately, Set 3 is the default for many education agencies today.

Please see the full paper for the traditional treatment of the issues. www.ARNIEdocs.info



Figure 3A: Body Mass Index Levels by Gender & Race/Ethnicity, 10 Super Heroes









Body Mass Index	Gender	African American	White	Hispanic	Japanese	Native American	Reptile	Total
Normal	Male	0			0		0	3
	Female	0		0		0	0	3
	Total	0	3	1	1	1	0	6
Overweight	Male			0	0	0	0	3
	Female	0	0	0	0	0	0	0
	Total	1	2	0	0	0	0	3
Obese	Male	0	0	0	0	0		1
	Female	0	0	0	0	0	0	0
	Total	0	0	0	0	0	1	1
Total	Male	1	3	1	0	1	1	7
	Female	0	2	0	1	0	0	3
	Total	1	5	1	1	1	1	10

Figure 3B: Body Mass Index Levels by Gender & Race/Ethnicity, Megateam of 100 Super Heroes

Body Mass Index	Gender	African American	White	Hispanic	Japanese	Native American	Reptile	Total
Normal	Male	0	10	10	0	10	0	30
	Female	0	20	0	10	0	0	30
	Total	0	30	10	10	10	0	60
Overweight	Male	10	20	0	0	0	0	30
	Female	0	0	0	0	0	0	0
	Total	10	20	0	0	0	0	30
Obese	Male	0	0	0	0	0	10	10
	Female	0	0	0	0	0	0	0
	Total	0	0	0	0	0	10	10
Total	Male	10	30	10	0	10	10	70
	Female	0	20	0	10	0	0	30
	Total	10	50	10	10	10	10	100

Figure 3C: Body Mass Index Levels by Gender & Race/Ethnicity, Set 3, De-Identified, 100 Super Heroes

Body Mass Index	Gender	African American	White	Hispanic	Japanese	Native American	Reptile	Total
Normal	Male	NA	NA	NA	NA	NA	NA	NA
	Female	NA	NA	NA	NA	NA	NA	NA
	Total	NA	NA	NA	NA	NA	NA	NA
Overweight	Male	NA	NA	NA	NA	NA	NA	NA
	Female	NA	NA	NA	NA	NA	NA	NA
	Total	NA	NA	NA	NA	NA	NA	NA
Obese	Male	NA	NA	NA	NA	NA	NA	NA
	Female	NA	NA	NA	NA	NA	NA	NA
	Total	NA	NA	NA	NA	NA	NA	NA
Total	Male	NA	NA	NA	NA	NA	NA	NA
	Female	NA	NA	NA	NA	NA	NA	NA
	Total	NA	NA	NA	NA	NA	NA	100

References

- (1) Louisiana Act No 837, 2014
- (2) The Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99. 1974
- (3) FERPA regulations. U.S. Department of Education:
www.ed.gov/policy/gen/reg/ferpa FERPA regulations
amendment. U.S. Department of Education (December
9, 2008):
www.ed.gov/legislation/FedRegister/finrule/2008-4/120908a.pdf
- (4) FERPA regulations amendment. U.S.
Department of Education (December 2, 2011):
www.gpo.gov/fdsys/pkg/FR-2011-12-02/pdf/2011-30683.pdf
- (5) Utah HB 358 Student Data Privacy Protection Act, 2016
- (6) Colorado HB 16-1423 Student Data Collection Use Security
- (7) Gurr, T. R. (1972). Politimetrics: An Introduction to
Quantitative Macropolitics. Englewood Cliffs, NJ: Prentice-
Hall.
- (8) Ligon, G. D. (2015) Demystifying De-Identification. Austin, Texas, ESP
Solutions Group, Inc.
- (9) IES/NCES. (2012). P-20W Data Governance: Tips from the
States. Statewide Longitudinal Data System Grant
Program.
- (10) Privacy Technical Assistance Center (PTAC), U.S.
Department of Education: <http://ptac.ed.gov>
- (11) Privacy Technical Assistance Center (Oct 2012): Case
Study #5: Minimizing Access to PII: Best
Practices for Access Controls and
Disclosure Avoidance Techniques.
<http://ptac.ed.gov/sites/default/files/case-study5-minimizing-PII-access.pdf>
- (12) Privacy Technical Assistance Center (Oct 2012):
Frequently Asked Questions—Disclosure Avoidance.
http://ptac.ed.gov/sites/default/files/FAQs_disclosure_avoidance.pdf
- (13) Statistical Policy Working Paper 22 - Report on Statistical Disclosure
Limitation Methodology.

Federal Committee on Statistical Methodology, Office of Management and Budget (1994):

<http://fcsm.gov/working-papers/wp22.html>

- (14) SLDS Technical Brief 1: Basic Concepts and Definitions for Privacy and Confidentiality in Student Education Records (NCES 2011-601):

<http://nces.ed.gov/pubs2011/2011601.pdf>

- (15) SLDS Technical Brief 3: Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting (NCES 2011-603):

<http://nces.ed.gov/pubs2011/2011603.pdf>

- (16) Technical Brief: Statistical Methods for Protecting Personally Identifiable Information in the Disclosure of Graduation Rates of First-Time, Full-Time Degree- or Certificate-Seeking Undergraduate Students by 2-Year Degree-Granting Institutions of Higher Education (NCES 2012-151): <http://nces.ed.gov/pubs2012/2012151.pdf>

- (17) Alaska Department of Education and Early Development. (2015) Alaska State Report Manager File Specifications.

- (18) Arizona Department of Education. (2015) Arizona Student Teacher Course Connection File Specifications, State Report Manager.

- (19) Connecticut Department of Education. (2015) Connecticut Vertical Reporting Framework File Specifications.

- (20) Iowa Department of Education. (2015) Iowa Vertical Reporting Framework File Specifications.

- (21) Idaho State Department of Education. (2015) Idaho State Report Manager File Specifications.

- (22) Missouri Department of Elementary and Secondary Education. (2015) MOSIS (Missouri Comprehensive Data System) Enterprise Metadata Dictionary, State Report Manager File Specifications.

- (23) Ohio Department of Education. (2015) Ohio Vertical Reporting Framework File Specifications.

- (24) Utah Department of Education. (2015) Vertical Reporting Framework File Specifications.

- (25) Wyoming Department of Education. (2015) WISE

- (Wyoming Integrated Statewide Education Data System)
Enterprise Metadata Dictionary, State Report Manager
File Specifications.
- (26) Alaska Department of Education and Early Development. (2015) Unity System.
 - (27) Colorado Department of Education. (2016) Enterprise Metadata Dictionary
 - (28) Delaware Department of Education. (2015) Delaware Insight Warehouse.
 - (29) Idaho State Department of Education. (2015) Idaho System for Education Excellence.
 - (30) Louisiana Department of Education. (2015) Louisiana Education Data Resource System.
 - (31) Missouri Department of Elementary and Secondary Education. (2015) Missouri Comprehensive Data System
 - (32) Montana State Department of Education. (2015) GEMS – Growth and Enhancement of Montana Students.
 - (33) South Dakota Department of Education. (2015) Student Teacher Accountability Reporting System (SD-STARS).
 - (34) Tennessee Department of Education. (2015) Tennessee Longitudinal Data System.
 - (35) Education Service Center Region 10. (2015) EDW - Education Data Warehouse.
 - (36) Wyoming Department of Education. (2015) Wyoming integrated Statewide Education Data System - WISE Data System.
 - (37) IES/NCES. Statewide Longitudinal Data System Grant Program. <https://nces.ed.gov/programs/slds/>
 - (38) NCES Privacy Technical Assistance Center. Data De-Identification: An Overview of Basic Terms. http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf
 - (39) Garfinkle, F. L. (2015). De-Identification of Personally Identifiable Information. National Institute of Standards and Technology. U.S. Department of Commerce.
 - (40) The Health Insurance Portability and Accountability Act (HIPAA) Pub.L. 104–191, 110 Stat. 1936. 1996
 - (41) U. S. Department of Health & Human Services. Guidance

Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

<http://www.hhs.gov/ocr/privacy/hipaa/understanding/coverentities/De-identification/guidance.html>

- (42) IES/NCES Weaving a Secure Web around Education.
https://nces.ed.gov/pubs2003/secureweb/ch_6.asp
- (43) PCI Security Standards Council. (2011) Standard: PCI Data Security Standard (PCI DSS) Version: 2.0 Author: Scoping SIG, Tokenization Taskforce.
- (44) Singh, Avinash C. (2010) "Maintaining Analytic Utility while Protecting Confidentiality of Survey and Nonsurvey Data," *Journal of Privacy and Confidentiality*: Vol. 1: Iss. 2, Article 3.
- (45) Raghunathan, T. E., Reither, J. P., and Rubin, D. B. Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics*, Vol. 19. No. 1, 2003, PP. 1-16.
- (46) Newton, I. (1687) *Mathematical Principles of Natural Philosophy*.
- (47) Vinten-Johansen, Peter (2003) *Cholera, Chloroform, and the Science of Medicine: A Life of John Snow*, Oxford University Press, Inc.





About ESP Solutions Group

ESP Solutions Group provides its clients with *Extraordinary Insight™* into P20W education data systems and analytics. 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 ESSA, SIF, Ed-Fi, *EDFacts*, CEDS, state reporting, metadata standards, data governance, data visualizations, and emerging 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 ESP Journal Series*, designed to help decision makers analyze, manage, and share data in the 21st Century.

Privacy versus Access to Data for School Improvement, Copyright © 2021 by ESP Solutions Group, Inc. All rights reserved. No part of this paper shall be reproduced, stored in a retrieval system, or transmitted by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the publisher.



ESP Solutions Group

(512) 879-5300

www.espsolutionsgroup.com