Protection of Personally Identifiable Information through Disclosure Avoidance Techniques

Michael Hawes
Statistical Privacy Advisor
U.S. Department of Education

Baron Rodriguez
Director
Privacy Technical Assistance Center

February 16, 2012

25th Annual Management Information Systems Conference

San Diego, CA

Presentation Overview

- Family Educational Rights and Privacy Act (FERPA)
- Disclosure Avoidance Primer
- ED's History with Disclosure Avoidance
- ED's Current Thinking
- Moving Forward
- Questions and Discussion

Family Educational Rights and Privacy Act (FERPA)

Definitions and Requirements

Confidentiality under FERPA

- Protects personally identifiable information (PII) from education records from unauthorized disclosure
- Requirement for written consent before sharing PII
- Exceptions from the consent requirement for:
  - "Studies"
  - "Audits and Evaluations"
  - Health and Safety emergencies
  - And others purposes as specified in §99.31

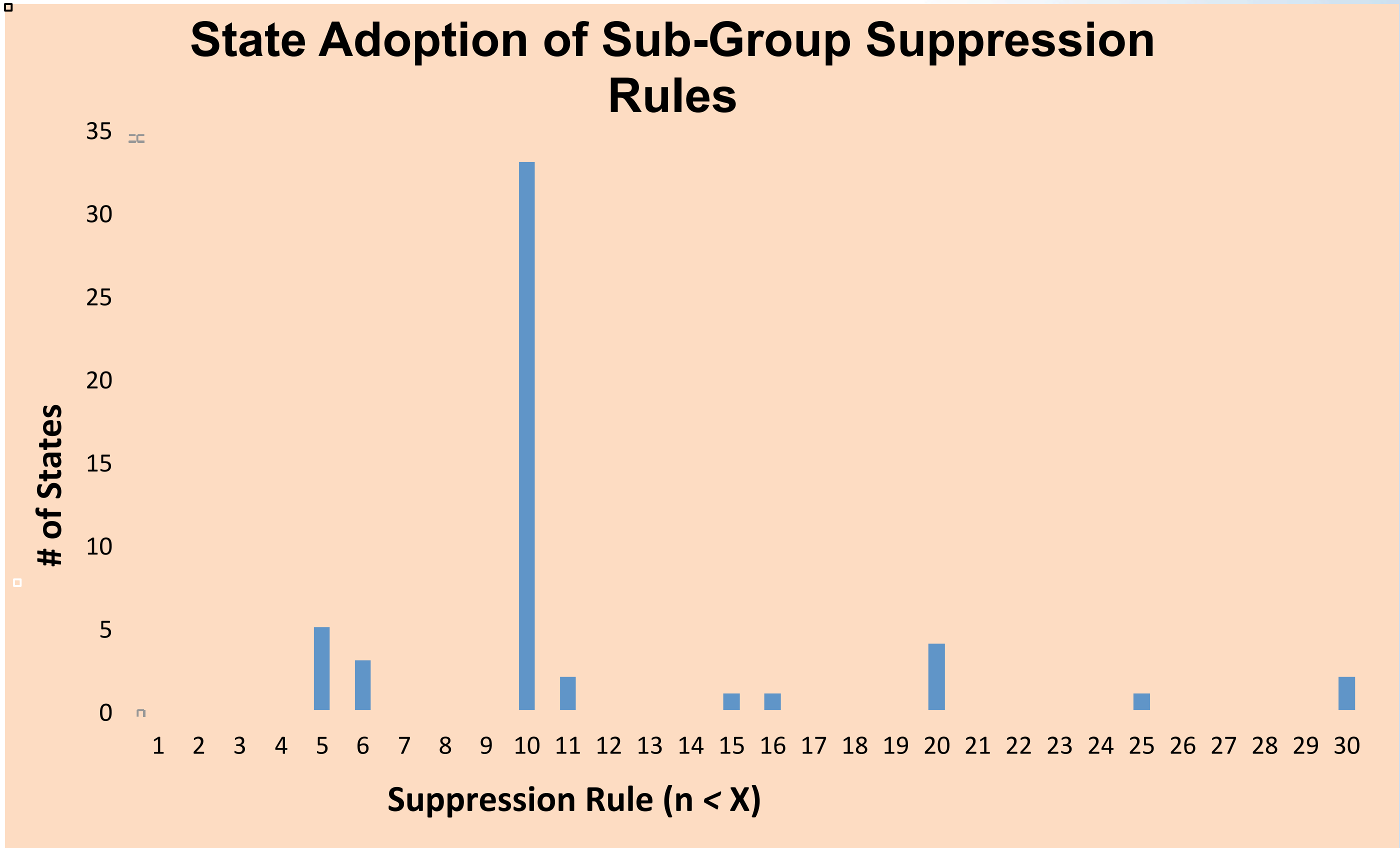
Personally Identifiable Information (PII)

- Name
- Name of parents or other family members
- Address
- Personal identifier (e.g., SSN, Student ID#)
- Other indirect identifiers (e.g., date or place of birth)
- *"Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty." (34 CFR § 99.3)*

Reporting vs. Privacy

- Department of Education regulations require reporting on a number of issues, often broken down across numerous sub-groups, including:
  - Gender
  - Race/Ethnicity
  - Disability Status
  - Limited English Proficiency
  - Migrant Status
  - Economically Disadvantaged Students
- BUT, slicing the data this many ways increases the risks of disclosure, and the regulations also require states to "*implement appropriate strategies to protect the privacy of individual students…*" (§200.7)

How States are Doing It



Example: School Performance Data

Sunshine Elementary 3rd Grade Class (Anywhere, U.S.A.)

	# Tested	Basic (and above)	Proficient (and above)	Advanced
Male	37	100%	59%	5%
Female	38	92%	66%	11%
AIAN	1	*	*	*
Black	37	92%	43%	5%
Hispanic	12	100%	75%	8%
Asian	4	*	*	*
White	21	100%	81%	5%
All Students	75	96%	63%	8%

Example: School Performance Data

Sunshine Elementary 3rd Grade Class (Anywhere, U.S.A.)

	# Tested	Basic (and above)	Proficient (and above)	Advanced
Male	37	(37) 100%	(22) 59%	(2) 5%
Female	38	(35) 92%	(25) 66%	(4) 11%
AIAN	1	*	*	*
Black	37	(34) 92%	(16) 43%	(2) 5%
Hispanic	12	(12) 100%	(9) 75%	(1) 8%
Asian	4	*	*	*
White	21	(21) 100%	(17) 81%	(1) 5%
All Students	75	(72) 96%	(47) 63%	(6) 8%

For each subgroup (row) multiply the percent by the # Tested to get the number of students in that category

Example: School Performance Data

Sunshine Elementary 3rd Grade Class (Anywhere, U.S.A.)

	# Tested	Basic (and above)	Proficient (and above)	Advanced
Male	37	(37) 100%	(22) 59%	(2) 5%
Female	38	(35) 92%	(25) 66%	(4) 11%
AIAN	1	(1) *	(1) *	(0-1) *
Black	37	(34) 92%	(16) 43%	(2) 5%
Hispanic	12	(12) 100%	(9) 75%	(1) 8%
Asian	4	(4) *	(4) *	(1-2) *
White	21	(21) 100%	(17) 81%	(1) 5%
All Students	75	(72) 96%	(47) 63%	(6) 8%

Calculate the suppressed subgroups by subtracting the remaining subgroup totals from the "All Students" totals

Example: School Performance Data

Sunshine Elementary 3rd Grade Class (Anywhere, U.S.A.)

	# Tested	Below Basic	Basic (and above)	Proficient (and above)	Advanced
Male	37	0	(37) 100%	(22) 59%	(2) 5%
Female	38	3	(35) 92%	(25) 66%	(4) 11%
AIAN	1	0	(1) *	(1) *	(0-1) *
Black	37	3	(34) 92%	(16) 43%	(2) 5%
Hispanic	12	0	(12) 100%	(9) 75%	(1) 8%
Asian	4	0	(4) *	(4) *	(1-2) *
White	21	0	(21) 100%	(17) 81%	(1) 5%
All Students	75	3	(72) 96%	(47) 63%	(6) 8%

Calculate the unreported outcome by subtracting the “Good” totals from the # Tested

But what is a disclosure anyway?

Disclosure Avoidance Primer

(aren’t you glad you had coffee this morning?)

It’s all about risk

“The release of any data usually entails at least some element of risk. A decision to eliminate all risk of disclosure would curtail [data] releases drastically, if not completely. Thus, for any proposed release of [data] the acceptability of the level of risk of disclosure must be evaluated.”

Federal Committee on Statistical Methodology, “Statistical Working Paper #2”

3 Basic Flavors of Disclosure Avoidance

- Suppression
- Blurring
- Perturbation

Suppression

Definition:	Removing data to prevent the identification of individuals in small cells or with unique characteristics
Examples:	• Cell Suppression • Row Suppression • Sampling
Effect on Data Utility:	• Results in very little data being produced for small populations • Requires suppression of additional, non-sensitive data (e.g., complementary suppression)
Residual Risk of Disclosure:	• Suppression can be difficult to perform correctly (especially for large multi-dimensional tables) • If additional data is available elsewhere, the suppressed data may be re-calculated.

Blurring

Definition:	Reducing the precision of data that is presented to reduce the certainty of identification.
Examples:	• Aggregation • Percents • Ranges • Top/Bottom-Coding • Rounding
Effect on Data Utility:	• Users cannot make inferences about small changes in the data • Reduces the ability to perform time-series or cross-case analysis
Residual Risk of Disclosure:	• Generally low risk, but if row/column totals are published (or available elsewhere) then it may be possible to calculate the actual values of sensitive cells

Perturbation

Definition:

Making small changes to the data to prevent identification of individuals from unique or rare characteristics

Examples:

- Data Swapping
- Noise
- Synthetic Data

Effect on Data Utility:

- Can minimize loss of utility compared to other methods
- Seen as inappropriate for program data because it reduces the transparency and credibility of the data, which can have enforcement and regulatory implications

Residual Risk of Disclosure:

- If someone has access to some (e.g., a single state’s) original data, they may be able to reverse-engineer the perturbation rules used to alter the rest of the data

The U.S. Department of Education's History with Disclosure Avoidance

How we got where we are today…

Recent Developments in Disclosure Avoidance at ED

- State Workbooks
- School and LEA level data
- Reactions from the field
- Technical Brief 3

ED's Current Thinking on Disclosure Avoidance

Emerging (but still unofficial) views

taking shape at ED

Emerging Views

- Perturbation and transparency
- Non-Trivial distinction between 0s and 1s
- Exceptions for publishing 100% in certain cases
- Who is a “reasonable person in the school community?”

Moving Forward?

Where do we go from here?

Moving Forward

- Data Release Working Group
- (Proposed) Formation of a Disclosure Review Board
- Guidance for the field

Our Goal: Publish as much usable data as we can AND protect privacy

Questions and Discussion

Baron Rodriguez Director Privacy Technical Assistance Center	Michael Hawes Statistical Privacy Advisor U.S. Department of Education
TEL: **(855) 249-3072**	TEL: **(202) 453-7017**
FAX: **(855) 249-3073**	FAX: **(202) 401-0920**
Email: PrivacyTA@ed.gov	Email: Michael.Hawes@ed.gov
Website: http://ptac.ed.gov	