



**FPF Student Privacy  
Train-the-Trainer Program  
Module 2: Defining Data  
CLE Materials**

*April 30, 2020*



## Module 2: Defining Data

April 30, 2020

### TABLE OF CONTENTS

<b>Module 2 Activities.....</b>	<b>4</b>
Understanding Types of Data.....	4
<i>Prior to the webinar, participants will read <a href="#">Case Study #5: Minimizing Access to PII: Best Practices for Access Controls and Disclosure Avoidance Techniques</a>, which explains direct identifiers, indirect identifiers, identifiable data, de-identified data, individual level data, aggregate data, and sensitive data. Then, they will plot each type of data on a graph with the x-axes labeled “individual level” and “aggregate” and the y-axes labeled “public” and “not public.”</i>	
Scenario: Informal Data Sharing.....	5
<i>Participants will read a scenario on informal data sharing and answer questions on how to develop, communicate, and enforce policies and procedures to prevent such disclosures in the future.</i>	
Scenario: Data Suppression Case.....	6
<i>Participants will consider a case where a state agency used data suppression in alignment with federal guidance documents but the public wants the unsuppressed document. As the state attorney, they will outline how they will build a case and explain the concept of indirect identifiers.</i>	
<b>Guide to De-Identified Data Handout.....</b>	<b>7</b>
<i>Note: This is original written material adapted from ongoing unpublished work that has been provided by Kelsey Finch <b>for this webinar only</b>.</i>	
<i>This handout explains the characteristics of identifiable data, contextual considerations for identifiable data, and legal considerations for identifiable data. It also compares and contrasts key laws explicitly defining direct identifiers, indirect identifiers, de-identification, and anonymization, including the California Consumer Protection Act (CCPA), the Children’s Online Privacy Protection Act (COPPA), the General Data Protection Regulation (GDPR), and the Health Insurance Portability and Accountability Act (HIPAA).</i>	
<b>Webinar Slide Deck.....</b>	<b>18</b>
<i>The webinar covers identifiers and de-identification core concepts in the context of the Family Educational Rights and Privacy Act (FERPA). It also describes steps to data de-identification and technical approaches to de-identification, particularly pseudonymization.</i>	

**Resources.....53**

A Visual Guide to Practical Data De-Identification [*Future of Privacy Forum*] .....53

Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting [*National Center for Education Statistics*] .....54

Ensuring Equity in ESSA: The Role of N-Size in Subgroup Accountability [*Alliance for Excellent Education*] .....85

The Importance of Disaggregating Student Data [*Safe Schools/Healthy Schools*] .....99

Understanding Differential Privacy and Why It Matters for Digital Rights [*Access Now*]......104

De-Identification and Student Data [*Future of Privacy Forum*] .....109

Note: CLE materials are provided digitally; attendees may receive a printed version of the materials upon request.



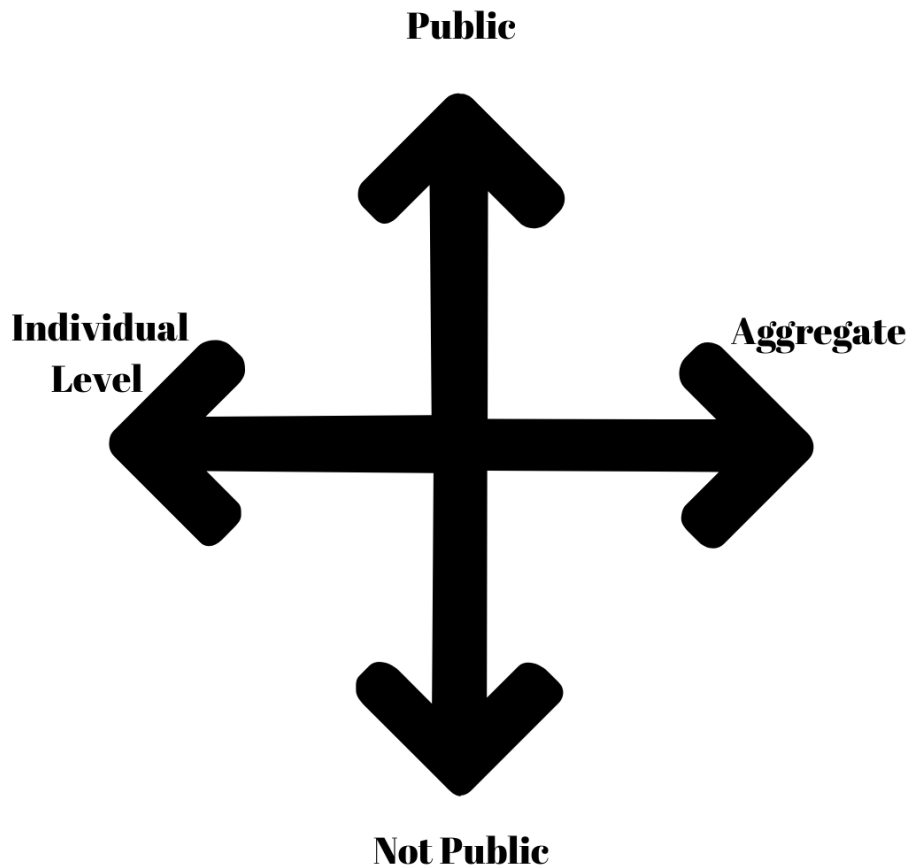
## Module 2 Activities

There are three individual activities for this module.

1. Case Study #5
2. Scenario: Informal Data Sharing
3. Scenario: Data Suppression Case

Please complete the activities in this document and email the document to [ttt@fpf.org](mailto:ttt@fpf.org) by **April 13th**.

1. Review [CASE STUDY #5: Minimizing Access to PII: Best Practices for Access Controls and Disclosure Avoidance Techniques](#). This study talks about direct identifiers, indirect identifiers, identifiable data, de-identified data, individual level data, aggregate data, and sensitive data. Plot each of these types of data on the graph below. If you are not able to print and scan the document, feel free to draw your own graph and upload a photo.





2. Review the scenario below and answer the following question.

**Scenario: Informal Data Sharing**

*“A school board member was a personal friend of the principal at the local elementary school. When the board member needed information, she would email the principal and get a reply with the data attached. Both school leaders knew they were circumventing official procedures for sharing data, but rationalized that, since they both had privileges to obtain the data from the data steward, this more direct and informal approach only expedited an exchange that was otherwise permissible anyway.*

*They didn't see any harm in this practice until the board member made a public presentation that inadvertently revealed that the one and only Asian female student in the 4th grade had a learning disability. The student's parents were in the audience and took offense to the public display of private information.”*

Excerpt from [NCES' Forum's Guide to Data Ethics](#).

**Question**

What measures should be adopted to prevent such disclosures in the future?

**Answer**



3. Review the scenario below and answer the following question.

**Scenario: Data Suppression Case**

**Question**

You have to defend a case where a state agency used suppression in alignment with federal guidance documents and the public wants the unsuppressed document. How do you build your case? How do you help the judge understand the concept of indirect identifiers?

**Answer**



## Module 2: Defining Data

*April 30, 2020*

*Note: This is original written material adapted from ongoing unpublished work that has been provided by Kelsey Finch **for this webinar only**.*

### Guide to De-Identified Data

At the heart of every privacy and data protection regulation is a question of identifiability. If a piece of information is personally identifiable, an array of privacy laws can be immediately brought to bear; however, if that same information can be rendered non-identifiable, or de-identified, the data might be “freed” from restrictions. Even when not required by law, minimizing the identifiability of data is an important security control and a recognized best practice for all organizations that hold personal information.

For privacy professionals, understanding when and how data crosses the threshold from identifiable to de-identified is critical. However, the legal standards and privacy enhancing technologies for de-identification are neither simple to understand nor easy to implement, and misapplying them can have significant consequences for organizations and individuals.

Determining when information crosses the threshold between identifiable and de-identified can be challenging. The goal of de-identification is to transform data in a way that protects privacy while maintaining as much analytic utility as possible. This guide is intended to help privacy professionals and their colleagues navigate the key contextual and legal considerations necessary to answer the question: **“How identifiable is this data?”**

Before applying any particular de-identification tool, organizations must understand the nature of their data. This guide consists of three sections, each dedicated to an important step in the process of determining identifiability. Privacy professionals should consult with technical, legal, and business experts within their organizations in using this guide, as appropriate.

- 1) Characteristics of Identifiable Data - understanding the basic characteristics of data, including spotting direct and indirect identifiers

- 2) Contextual Considerations for Identifiable Data - understanding important technical and environmental factors that impact identifiability
- 3) Legal Considerations for Identifiable Data - understanding different legal standards of identifiability

We hope that this resource will provide privacy professionals and their colleagues a starting point and common language for navigating de-identification principles and practices. By conducting this initial analysis, organizations will be better able to identify the combination of technical methodologies and organizational controls that best fits their data and circumstance.

*De-identification tools must be applied and assessed on a case-by-case basis. This guide is a navigational tool, and does not constitute legal advice.*

### **Step One: Characteristics of Identifiable Data**

The first step to assessing whether data is identifying, identifiable, or de-identified is understanding its basic characteristics. Here, factors such as whether the data contains direct or indirect identifiers and whether it is in a particular format are considered. These factors impact the data's identifiability and what measures an organization must take to de-identify it.

[1.a Is there a direct identifier?](#) Data that contains a direct identifier is identifying. In order to be considered de-identified, all direct identifiers must be eliminated or transformed.

[1.b Is there an indirect identifier?](#) Data that contains an indirect identifier is identifiable. In order to be considered de-identified, all direct *and* indirect identifiers must be eliminated or transformed.

[1.c What format is this data?](#) Data that is in an unstructured or dynamic format may be identifiable, and requires special attention and tools. In order to be considered de-identified, all direct *and* indirect identifiers must be eliminated or transformed.

### **Step Two: Contextual Considerations for Identifiable Data**

Measuring the technical probability of re-identification depends on a variety of factors, including: the nature of the original data, the technical skill and resources of the "attacker," and the availability of additional information that can be linked with the de-identified data. These factors are context-specific, and in practice should be quantified on a case-by-case basis by a qualified expert. Additionally, some types of data have inherent characteristics -- including their uniqueness, persistence, and prevalence -- that impact their identifiability.



[2.a Outside Data Availability: How linkable is this information to other data \(now and in the future\)?](#)

[2.b Recipients' Resources: How sophisticated are potential "attackers"?](#)

[2.c Prevalence: How widely-used or how common is the identifier?](#)

[2.d Persistence: Is the identifier persistent or dynamic?](#)

[2.e Uniqueness: How many individuals are tied to the identifier?](#)

[2.f Data Sharing: Will the data be shared publicly?](#)

### **Step Three: Legal Considerations for Identifiable Data**

The final step in assessing identifiability is considering the legal context. While privacy laws do not always keep pace with technical possibilities, including around de- and re-identification standards -- it is important for privacy professionals to address both. Policymakers have taken a variety of approaches to defining identifiable data, and databases with similar technical characteristics could be treated differently from one jurisdiction to another.

[3.a Role of De-identification: Is it a process or as an outcome?](#)

[3.b Controls and safeguards: Do administrative and legal controls count?](#)

[3.c Pseudonymization: How is pseudonymous data treated?](#)

[3.d Disclosure Risks: What kind of disclosures are being protected?](#)

### **Conclusion**

Understanding the identifiability of personal data is a critical skill for privacy professionals. Of course, identifiability is only *one* factor in assessing and mitigating privacy risk; data that is marginally identifiable may nevertheless be highly consequential to individuals. Other factors like the sensitivity of data, the vulnerability of data subjects, and organizations' ethical responsibilities must also be taken into account when determining how data will be responsibly collected, used, and safeguarded.

**Appendix of Key Laws (a) Explicitly Defining Direct & Indirect Identifiers, (b) Explicitly Mentioning/Defining De-Identification or Anonymization, and (c) Enumerating Standards for Removing “Linkability”**

*Key Laws Explicitly Defining Direct and Indirect Identifiers:*

<b>CCPA</b>	
<b>Direct Identifiers</b>	
Name	X - real name; alias
ID number	X - unique personal identifier, SSN
E-mail Address	X
Government-issued identifiers	X - SSN, driver’s license number, passport number; state identification card number
Telephone Number	X
Physical Address	X - postal address; address
Biometric identifiers (e.g., fingerprint, voice print)	X - imagery of the iris, retina, fingerprint, face, hand, palm, vein patterns, and voice recordings, from which an identifier template, such as a faceprint, a minutiae template, or a voiceprint, can be extracted, and keystroke patterns or rhythms, gait patterns or rhythms, and sleep, health, or exercise data that contain identifying information.
Financial account information	X - bank account number, credit card number, debit card number,
Other direct identifiers	X - insurance policy number
<b>Indirect Identifiers</b>	

Persistent Identifiers	X - Persistent identifiers that can be used to identify a particular consumer or device, including but not limited to: a unique personal identifier, online identifier, IP address, account name, device identifier, cookies, beacons, pixel tags, mobile ad identifiers, or similar technology; customer number, unique pseudonym, or user alias.
Probabilistic IDs	X - Probabilistic identifiers that can be used to identify a particular consumer or device, including but not limited to: (same as directly above)

COPPA	
Direct Identifiers	
Name	X - first and last name
ID number	X - SSN
E-mail Address	X - online contact information, including a screen or user name that functions as online contact information
Government-issued identifiers	X - SSN
Telephone Number	X
Physical Address	X - home or other physical address including street name and name of a city or town
Biometric identifiers (e.g., fingerprint, voice print)	X - a photograph, video, or audio file, where such file contains a child's image or voice
Indirect Identifiers	
Persistent Identifiers	X - a persistent identifier that can be used to recognize a user over time and across different websites or online services

Probabilistic IDs	X - Geolocation information sufficient to identify street name and name of a city or town
-------------------	---

FERPA	
Direct Identifiers	
Name	X
ID number	X - SSN; Student ID
Physical Address	X - address of the student or student's family
Biometric identifiers (e.g., fingerprint, voice print)	X - biometric record
Indirect Identifiers	
Date of birth	X
Place of Birth	X
Mother's Maiden Name	X
Probabilistic IDs	X - Name of the student's parent or other family members

GDPR	
Direct Identifiers	
Name	X
ID number	X
Biometric identifiers (e.g., fingerprint, voice print)	X - an identifier consisting of “one or more factors specific to” the “physical,” “physiological,” or “genetic” identity of a natural person.

Financial account information	X - an identifier consisting of “one or more factors specific to” the “economic” identity of a natural person.
<b>Indirect Identifiers</b>	
Persistent Identifiers	X - online identifier, internet protocol addresses, cookie identifiers or other identifiers such as radio frequency identification tags.

<b>HIPAA</b>	
<b>Direct Identifiers</b>	
Name	X
ID number	X - SSN; account numbers
E-mail Address	X
Government-issued identifiers	X - SSN; license plate numbers
Telephone Number	X
Physical Address	X - address
Biometric identifiers (e.g., fingerprint, voice print)	X - Biometric identifiers, including finger and voice prints; Full face photographic images and any comparable images
Other direct identifiers	X - Medical record numbers; Health plan beneficiary; fax numbers; vehicle identifiers and serial numbers;
<b>Indirect Identifiers</b>	
Date of birth	X
Age	X

ZIP Code	X
Persistent Identifiers	X - Device identifiers and serial numbers; IP address numbers

*Key Laws **Explicitly** Mentioning/Defining De-identification or Anonymization:*

Law	Content	Citation
<b>CCPA</b>	<p><b>“Deidentified”</b> means information that cannot reasonably identify, relate to, describe, be capable of being associated with, or be linked, directly or indirectly, to a particular consumer, provided that a business that uses de-identified information:</p> <ol style="list-style-type: none"> <li>1. Has implemented technical safeguards that prohibit reidentification of the consumer to whom the information may pertain</li> <li>2. Has implemented business processes that specifically prohibit reidentification of the information.</li> <li>3. Has implemented business processes to prevent inadvertent release of de-identified information.</li> <li>4. Makes no attempt to re-identify the information.</li> </ol> <p>“Personal information” does not include consumer information that is de-identified or aggregate consumer information.</p>	1798.140 (h)(1)-(4); 1798.140 (o)(3)
<b>HIPAA</b>	<p>HIPAA has no restrictions on the use or disclosure of <b>de-identified</b> health information. De-identified health information neither identifies nor provides a reasonable basis to identify an individual. There are two ways to de-identify data: (1) determination by a qualified statistician or (2) removing the various identifiers found in 45 C.F.R. §164.514 (b)(2)(i)(A-K), and ensuring that the covered entity does not have actual knowledge that the information disclosed could be used alone or in combination with other information to identify the HIPAA protected individual.</p>	45 C.F.R 164.514(a); 45 C.F.R.514(b)(2)

<b>FTC</b>	To determine when data are not “reasonably linkable,” the FTC has established a Three-Part Test. According to the Test, data is not “reasonably linkable” to a particular consumer or device to the extent that a company: (1) <b>takes reasonable measures to ensure that the data are de-identified</b> ; (2) publicly commits not to try to re-identify the data; and (3) contractually prohibits downstream recipients from trying to re-identify the data.	<a href="#">Protecting Consumer Privacy in an Era of Rapid Change</a> , p.iv.
------------	---	---

*Key Laws Enumerating Standards for Removing “Linkability”:*

<b>Law</b>	<b>Content</b>	<b>Citation</b>
FTC	<p>To determine when data are not “reasonably linkable,” the FTC has established the Three-Part Test. According to the Test, data is not “reasonably linkable” to individual identity to the extent that a company: (1) takes “reasonable measures” to ensure that the data are de-identified; (2) publicly commits not to try to re-identify the data; and (3) contractually prohibits downstream recipients from trying to re-identify the data.</p> <p>“Reasonable measures” require that a company “achieve a reasonable level of justified confidence that the data cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer, computer, or other device.” Determining what qualifies as a “reasonable” level of justified confidence is circumstantial, depending in part on the available methods and technologies, the nature of the data at issue, and the purposes for which it will be used.</p>	<a href="#">Protecting Consumer Privacy in an Era of Rapid Change</a> , p.iv & 21.
HIPAA	HIPAA provides that organizations may deem health data “ <b>de-identified</b> ” using the “safe harbor” method, by which eighteen categories of identifiers are removed from a data file, after which data can be released publicly. Such data can include a special purpose code of identification allowing the organization that created the data to re-identify individuals, as	45 C.F.R. § 164.514(a) and (b)

	<p>long as the identifier is not related to information about the individual and cannot be used by others to identify the individual. If the data is shared under contractual protections for limited research, public health, or health care operations, the data may include specific dates and other indirect identifiers. But in neither case can an IP address be included. Another way to de-identify data by a formal determination by a qualified statistician.</p>	
GDPR	<p>“To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly.” To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective <i>factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.</i></p>	Recital 26

### Additional Resources and Implementation Guidance

- **UK Anon, Anonymisation Decision-Making Framework** - <http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf>
- **NIST, De-Identification of Personal Data** - <http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf>
- **NIST, De-Identifying Government Datasets** - [http://csrc.nist.gov/publications/drafts/800-188/sp800\\_188\\_draft.pdf](http://csrc.nist.gov/publications/drafts/800-188/sp800_188_draft.pdf)
- **De-Identification Maturity Model** - [https://iapp.org/media/pdf/resource\\_center/2014-14-05%20Privacy%20Analytics%20The%20De-identification%20Maturity%20Model.pdf](https://iapp.org/media/pdf/resource_center/2014-14-05%20Privacy%20Analytics%20The%20De-identification%20Maturity%20Model.pdf)
- **HITRUST De-Identification Framework** - <https://hitrustalliance.net/de-identification/>
- **Dep’t of Ed/PTAC, Basic Terms Overview** - [https://studentprivacy.ed.gov/sites/default/files/resource\\_document/file/data\\_deidentification\\_terms.pdf](https://studentprivacy.ed.gov/sites/default/files/resource_document/file/data_deidentification_terms.pdf)
- **Article 29 WP Opinion on Anonymisation Techniques** - [https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216\\_en.pdf](https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf)



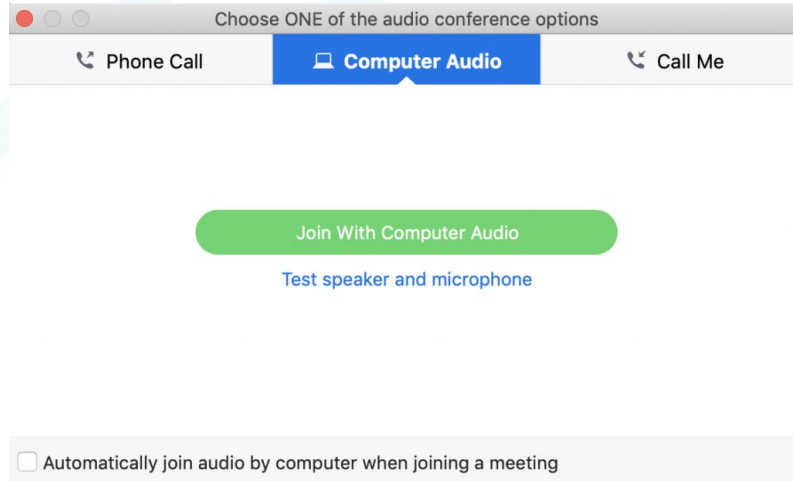
- Berkman Klein Center, Towards a Modern Approach to Privacy-Aware Government Data Releases - [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2779266](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2779266)



# **TRAIN** THE **TRAINER**

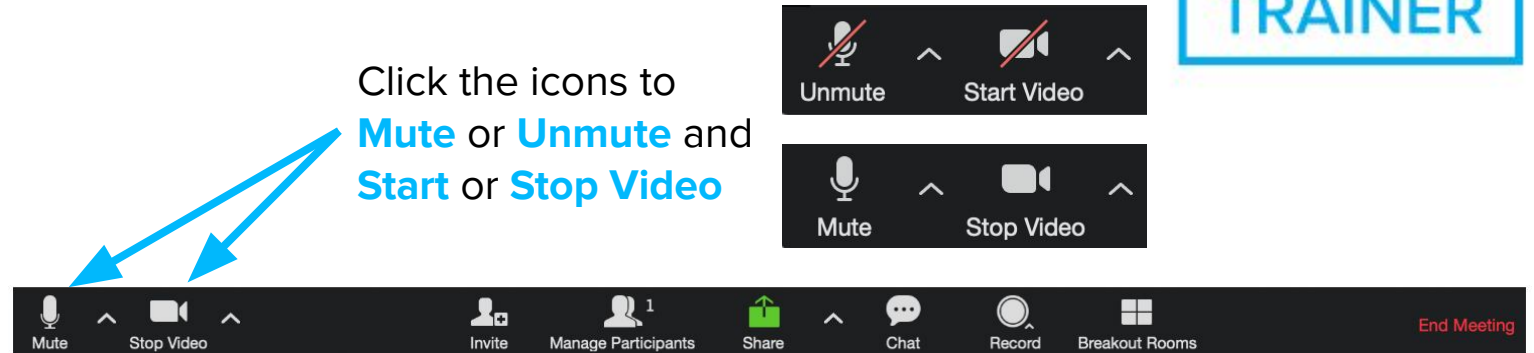
## **MODULE 2 WEBINAR**

# ZOOM BASICS



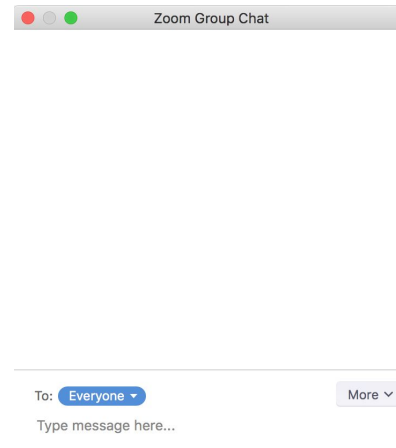
You can **Join by Computer Audio** using your Microphone or **Join by Phone**. You also have the option to Test your speaker and microphone.

Click the icons to **Mute** or **Unmute** and **Start** or **Stop Video**



## Conduct a group or private chat

Click on **Chat**. Type a message and hit “enter” on your keyboard to send a chat to Everyone. You can also message an individual participant via private chat by clicking on the drop down for **To:** and selecting an individual’s name.



# ZOOM AUDIO/VIDEO ISSUES



## AUDIO ISSUES

**Can't hear the other participants or other participants can't hear you?**

Make sure your computer speaker volume is turned up

Make sure your speakers are selected for active output in Zoom

Make sure your microphone is selected for active input in Zoom

- In the Zoom meeting, click the **up arrow next to Audio** and select **Test Computer Mic & Speakers**
- Click the **Test Speaker** button. If you hear audio, this is setup correctly. If you do not hear audio, use the drop down box and select a different output and press Test Speaker again.
- Click the **Test Mic** button. If you see green bars in the volume meter when you speak, this is setup correctly. If you do not see the green volume meter bars or hear the audio message you recorded, use the drop down box and select another mic and press Test Mic again.

## VIDEO ISSUES

**Can't see the other participants or other participants can't see you?**

Make sure you have installed the Zoom software and are logged into the meeting.

Make sure your camera is turned on, plugged in, and selected in Zoom.

Make sure your camera is selected in the video section.

- In the Zoom meeting, click the **up arrow next to Video** and select your preferred camera under **Select a Camera**.

# AGENDA



1. A Quick, Anonymous, Ungraded Quiz
2. Module 2 Objectives
3. Module 2 Activities
4. Presentation
5. Q&A

# QUIZ



1. Which term describes records that have enough PII removed or obscured so that the remaining information does not identify an individual and there is no reasonable basis to believe that the information can be used to identify an individual?
  - a. Directory Information
  - b. De-Identified Data
  - c. Encrypted Data

# QUIZ



1. Which term describes records that have enough PII removed or obscured so that the remaining information does not identify an individual and there is no reasonable basis to believe that the information can be used to identify an individual?
  - a. Directory Information
  - b. De-Identified Data
  - c. Encrypted Data

# QUIZ



2. Which of the following is **NOT** an example of a disclosure limitation method?
- a. Perturbation
  - b. Suppression
  - c. Transforming
  - d. Blurring
  - e. Coarsening
  - f. Masking



# QUIZ



2. Which of the following is NOT an example of a disclosure limitation method?
- a. Perturbation
  - b. Suppression
  - c. Transforming
  - d. Blurring
  - e. Coarsening
  - f. Masking

# QUIZ



3. What is removing data from a cell or row in a table to prevent the identification of individuals in small groups or those with unique characteristics an example of?
- a. Perturbation
  - b. Suppression
  - c. Transforming
  - d. Blurring
  - e. Coarsening
  - f. Masking

# QUIZ



3. What is removing data from a cell or row in a table to prevent the identification of individuals in small groups or those with unique characteristics an example of?
- a. Perturbation
  - b. Suppression**
  - c. Transforming
  - d. Blurring
  - e. Coarsening
  - f. Masking

# PTAC GLOSSARY



More definitions here: <https://studentprivacy.ed.gov/glossary>

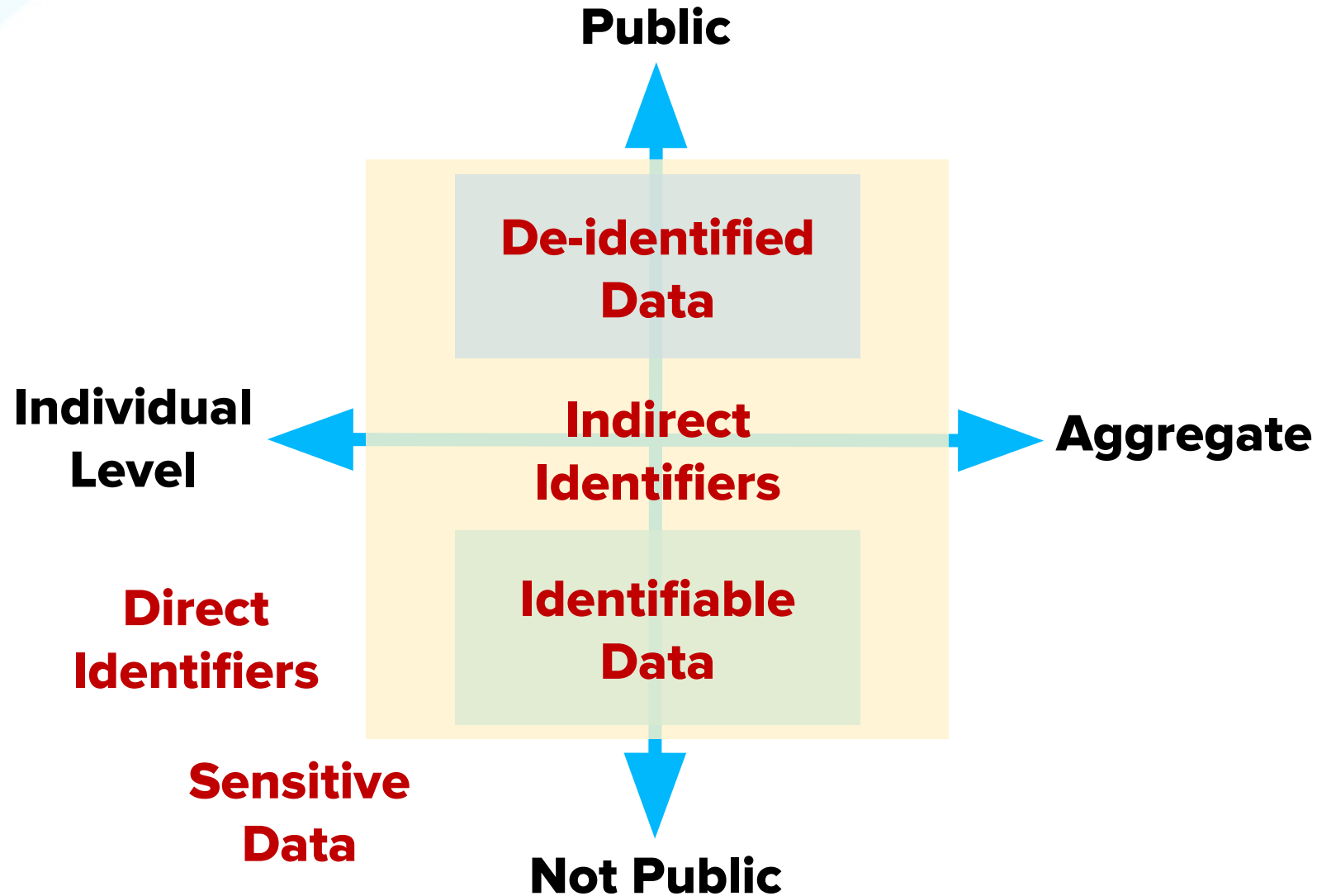
- **Blurring:** a disclosure limitation method which is used to reduce the precision of the disclosed data to minimize the certainty of individual identification
- **Coarsening:** a disclosure limitation method which preserves the individual respondent's data by reducing the level of detail used to report some variables
- **De-identified Data:** records that have a re-identification code and have enough personally identifiable information removed or obscured so that the remaining information does not identify an individual and there is no reasonable basis to believe that the information can be used to identify an individual
- **Masking:** a disclosure limitation method that is used to “mask” the original values in a data set to achieve data privacy protection
- **Perturbation:** a disclosure limitation method which involves making small changes to the data to prevent identification of individuals from unique or rare population groups
- **Suppression:** a disclosure limitation method which involves removing data (e.g., from a cell or a row in a table) to prevent the identification of individuals in small groups or those with unique characteristics

# MODULE 2 OBJECTIVES



1. Clarify the meaning of **identifiable** and **de-identified** data.
2. Compare and contrast **individual level** and **aggregate level** data.
3. Use and defend **statistical suppression**.

# MODULE 2 ACTIVITIES



# MODULE 2 ACTIVITIES



## **SCENARIO: Informal Data Sharing**

*“A school board member was a personal friend of the principal at the local elementary school. When the board member needed information, she would email the principal and get a reply with the data attached. Both school leaders knew they were circumventing official procedures for sharing data, but rationalized that, since they both had privileges to obtain the data from the data steward, this more direct and informal approach only expedited an exchange that was otherwise permissible anyway.*

*They didn't see any harm in this practice until the board member made a public presentation that inadvertently revealed that the one and only Asian female student in the 4th grade had a learning disability. The student's parents were in the audience and took offense to the public display of private information.”*

Excerpt from [NCES' Forum's Guide to Data Ethics](#)

# MODULE 2 ACTIVITIES



## **SCENARIO: Informal Data Sharing**

### **Preventative Measures from Your Answers**

- Develop policies for unintentional (or intentional) disclosures procedures and consequences for breaking protocol
- Review data sharing policies and procedures with school leadership and school board
- Enforce or develop policies for unintentional (or intentional) disclosures
- Provide training to district staff and school board members, explaining why these procedures are important, and inform of possible consequences
- All data requests should be processed through the district's Data Steward to allow for comprehensive review, discussion, and planning, as necessary, before approval



# MODULE 2 ACTIVITIES



## **SCENARIO: Data Suppression Case**

*You have to defend a case where a state agency used suppression in alignment with federal guidance documents and the public wants the unsuppressed document. How do you build your case? How do you help the judge understand the concept of indirect identifiers?*

# PRESENTATION



**Kelsey Finch**  
*Senior Counsel*  
*Future of Privacy Forum*



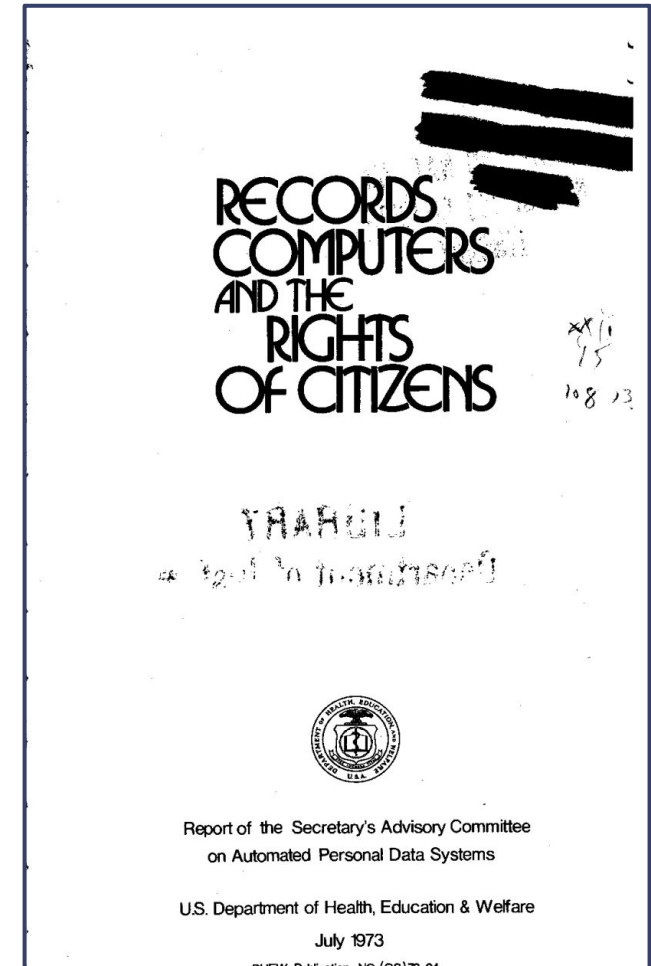
**David Rubin**  
*Attorney at Law*  
*David B. Rubin P.C.*



**TRAIN** THE  
**TRAINER**

# Rise of “Personal Information” as a Central Concept in U.S. Privacy Law

- Before computers vs. after the computer revolution (late 1960's)
- Fair Information Practices (FIPs)
- Evolving legislative strategies
  - Early privacy laws focused on treatment of “records” about people and the management of recordkeeping systems - e.g. FCRA 1970, FERPA 1974, Privacy Act 1974
  - Shift towards regulating “personal information”
    - e.g., The Cable Act 1984
    - Since then, U.S. privacy laws have continued to use the collection of personal information as the trigger for the applicability of legal protections
    - **Ongoing problems** often stem from lack of consensus on a single concept of “personal information”



# The Advent of “Big Data”

- 2005 Grants for Statewide Longitudinal Data Systems
- 2006 Facebook available to the public
- 2007 The first iPhone, Twitter spins off as its own company, Google launches Android, GitHub, IBM begins building Watson
- 2008 “School official” exception (34 CFR § 99.31(a)(1)(i)).
- 2009 SLDS mandatory for “Race to the Top” funds

# School District Privacy Concerns

Respecting student privacy for its own sake

Monetization of data

Marketing to students

Threats to student safety (2018 FBI Alert)

Unfair profiling of students with long-term implications

# Challenges for School Attorneys

FERPA adopted in the “Stone Age”

Student data privacy often relegated to “tech” staff alone

Lack of technical sophistication among school districts and their attorneys

Limited legal exposure (Gonzaga University v. Doe (USSC 2002))

Changing mindset of district staff to think in terms of “data” not “records”



# A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.



**DEGREES OF IDENTIFIABILITY**  
Information containing direct and indirect identifiers.



**PSEUDONYMOUS DATA**  
Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.

































**DE-IDENTIFIED DATA**  
Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.



**ANONYMOUS DATA**  
Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

This is a primer on how to distinguish different categories of data.

- **DIRECT IDENTIFIERS**  
Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)
- **INDIRECT IDENTIFIERS**  
Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)
- **SAFEGUARDS and CONTROLS**  
Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals

EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
 INTACT	 PARTIALLY MASKED	 PARTIALLY MASKED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED
 INTACT	 INTACT	 INTACT	 INTACT	 INTACT	 INTACT	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED
 NOT RELEVANT due to nature of data	 LIMITED or NONE IN PLACE	 CONTROLS IN PLACE	 CONTROLS IN PLACE	 LIMITED or NONE IN PLACE	 CONTROLS IN PLACE	 LIMITED or NONE IN PLACE	 CONTROLS IN PLACE	 NOT RELEVANT due to nature of data	 NOT RELEVANT due to high degree of data aggregation
<b>SELECTED EXAMPLES</b> Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = SL7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)	Same as De-identified, except data are also protected by safeguards and controls	For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)	Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)



# Basics of Identifiers

## Direct identifiers:



## Indirect identifiers:



## Sensitive data:





# Identifiers in Context: FERPA

**Personally identifiable information** includes, but not limited to:

- Names (student, parent or family member; address (student or family); a personal identifier such as SSN, student # or biometric record; other indirect identifiers (DOB, place of birth, mother's maiden name)

**De-identified/anonymous data:** Not explicitly defined.

# De-Identification Core Concepts: What Is Protected

## *Which disclosure risk(s)?*

- ▶ Identity disclosed
- ▶ Contactable
- ▶ Singled out, treated differently
- ▶ Related, connected, or linked to
- ▶ Something new revealed
- ▶ Individuals vs. groups

## *Which “attackers”?*

- ▶ General public/ordinary people
- ▶ Experts
- ▶ Nosy neighbors (insider knowledge)
- ▶ Journalists, prosecutors, grad students\* (motivated intruders)
- ▶ Criminals
- ▶ Data brokers

## *What context?*

- ▶ Confidence threshold
- ▶ The same data might be personal in one context, but de-identified in another (or in the hands of another organization)
- ▶ Data about one person may also be data about others (shared devices, households, genetics)

## *How suited to standard de-identification tools?*

- ▶ Contact information (name, phone, email, SSN)
- ▶ Biometrics and genetic information
- ▶ Precise geolocation
- ▶ Images, video, audio
- ▶ Unstructured
- ▶ Outliers

# Core Concepts in Context: FERPA

**Personally identifiable information** includes, but not limited to:

- Names (student, parent or family member; address (student or family); a personal identifier such as SSN, student # or biometric record; other indirect identifiers (DOB, place of birth, mother's maiden name)
- 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**
- Or information requested by a person who the educational agency or institution **reasonably believes knows** the **identity** of the student to whom the education record relates

**Process for releasing de-identified data:**

- removal of all personally identifiable information, provided that the education agency...has made a **reasonable determination** that a student's identity is not personally identifiable, whether through single or multiple releases, and taking into account **other reasonably available information**

L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

## Simple Demographics Often Identify People Uniquely

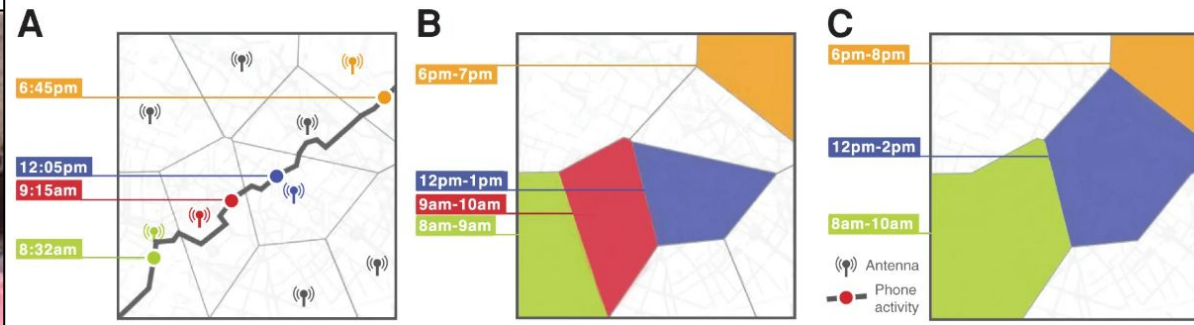
## Robust De-anonymization of Large Sparse Datasets

NETFLIX

Netflix Prize

Arvind Narayanan and Vitaly Shmatikov  
The University of Texas at Austin

From: Unique in the Crowd: The privacy bounds of human mobility



## The New York Times

### *The Golden State Killer Is Tracked Through a Thicket of DNA, and Experts Shudder*

### Public NYC Taxicab Database Lets You See How Celebrities Tip



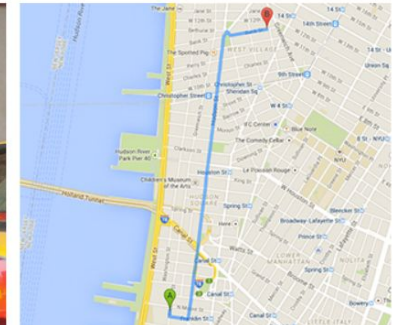
J.K. Trotter

10/23/14 12:00PM Filed to: DATA

142.43K



BRADLEY COOPER



JULY 8, 2013 • 7:34 PM - 7:44 PM  
376 GREENWICH ST. TO 13 BANK ST.  
\$9.00 FARE • CASH; UNKNOWN TIP • ©SPLASH

Fitness tracking map reveals U.S. bases

Kandahar airbase, Afghanistan

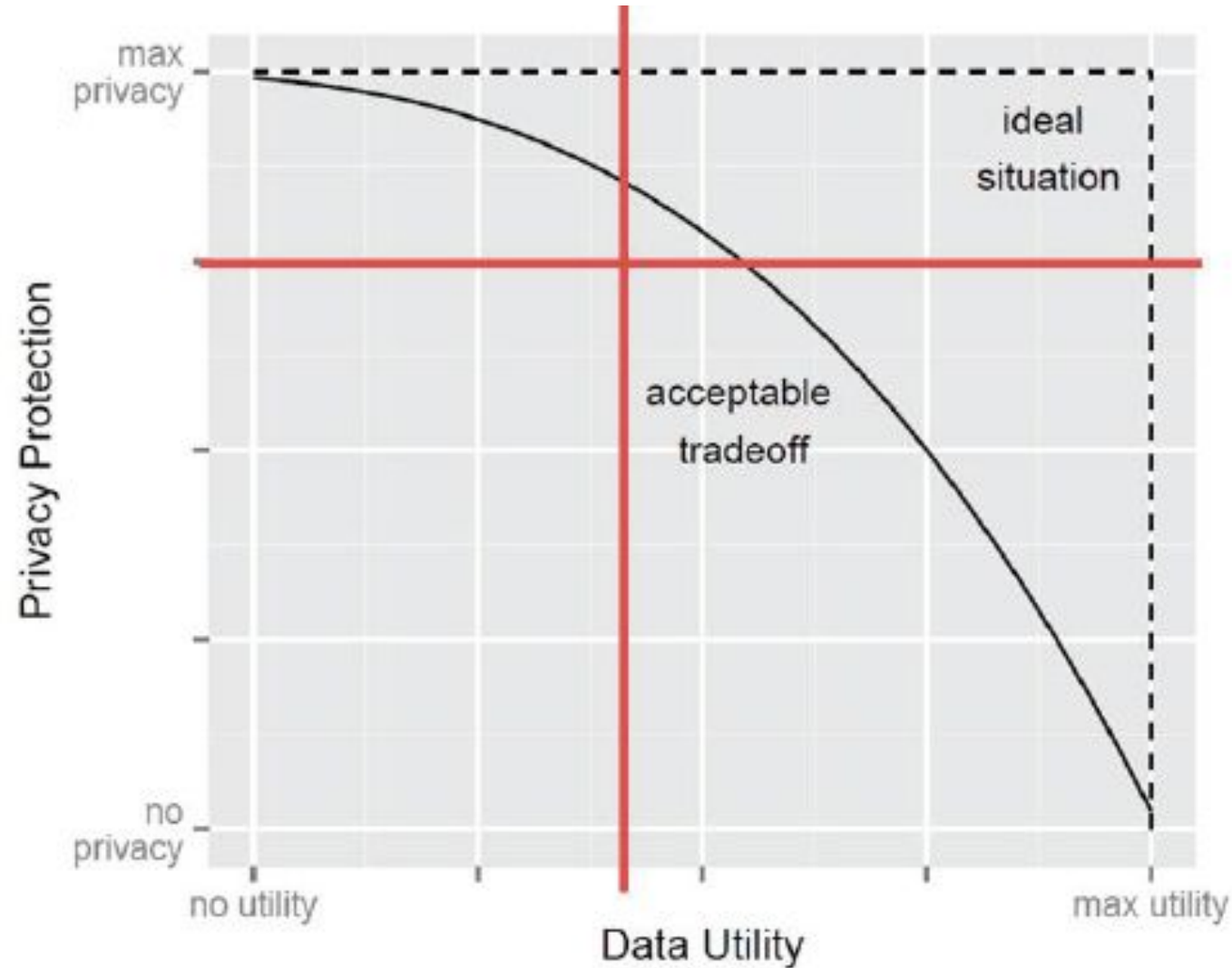


GPS tracking company Strava published an interactive map in Nov. 2017, showing where people have used fitness tracking devices. (Patrick Martin/The Washington Post)

TECHNOLOGY

*A Face Is Exposed for AOL Searcher No. 4417749*

# Core Challenge: Privacy vs. Utility





# Steps to Data De-Identification

1. **Determine your privacy, data usability, and access objectives** (what data quality, re-id risk, resources, release model are acceptable for this situation?)
2. **Conduct a data survey** (direct and indirect IDs? specialty data types/formats?)
3. **Apply technical treatments**
  - Traditional statistical disclosure limitation methods (remove direct IDs, transform indirect IDs)
  - Emerging and formal methods (differential privacy, synthetic data)
4. **Validate the de-identified dataset** (usefulness; privacy protection – “motivated intruder” test)
5. **Data release** (release & forget; data use agreements; secure enclave)
6. **Post-release monitoring**

# Technical Approaches to De-ID

## *Traditional methods: statistical disclosure controls*

- ▶ Suppression
- ▶ Blurring/generalization
- ▶ Perturbation (adding random noise)
- ▶ Aggregation

## *Legal & organizational controls*

- ▶ Contracts and use limitations
- ▶ Prohibitions on re-identification
- ▶ Access and security controls
- ▶ Ethical and disclosure review boards

## *Emerging techniques*

- ▶ Differential privacy
- ▶ Secure multi-party computation (including fully homomorphic encryption)
- ▶ Synthetic data



# Pseudonymization

Pseudonymization can refer to either:

- (1) a **legal category** of identifiable but not identified data, and
- (2) a **de-identification technique**



## Why organizations pseudonymize data:

- ▶ To optimize utility/privacy tradeoff, enabling a wider range of productive uses of data
- ▶ Reduces but doesn't eliminate compliance obligations; data stay "in scope" of privacy law

## Challenges of pseudonymous data:

- ▶ What is identifiable can be a moving target
- ▶ Requires a mix of technical and legal safeguards
- ▶ Should not be made public

- ▶ Emerging laws: device and probabilistic IDs in state legislation (CCPA)
- ▶ Existing laws: HIPAA Limited Data Sets; **FERPA research exemptions**; GDPR
- ▶ As technique: when directly identifying information is replaced with a new code or pseudonym ("Kelsey Finch" → "User 27399CA12").



# De-Identification in Summary

## De-identification is *not*:

- ▶ A Silver Bullet
- ▶ One-size-fits-all
- ▶ A set-it-and-forget safeguard; real resources and re-evaluation are frequently required

## De-identification *is* useful for:

- ▶ Mitigating privacy risks in a variety of scenarios
- ▶ Data sharing
- ▶ Data security
- ▶ “Privacy by Design” and Privacy by default
- ▶ Enabling socially beneficial research & data uses (medical, social sciences, economic, statistical, etc.)

# FPF Resources



## De-Identification & Student Data

August 2015

Reg Leichy  
Foresight Law + Policy  
Brenda Leong  
Future of Privacy Forum

### A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.

This is a primer on how to distinguish different categories of data.



**DEGREES OF IDENTIFIABILITY**  
Information containing direct and indirect identifiers.



**PSEUDONYMOUS DATA**  
Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.



**DE-IDENTIFIED DATA**  
Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.



**ANONYMOUS DATA**  
Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

- DIRECT IDENTIFIERS**  
Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)
- INDIRECT IDENTIFIERS**  
Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)
- SAFEGUARDS and CONTROLS**  
Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals

EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
INTACT	PARTIALLY MASKED	PARTIALLY MASKED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
NOT RELEVANT due to issue of data	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	NOT RELEVANT due to high degree of data aggregation	NOT RELEVANT due to high degree of data aggregation

#### SELECTED EXAMPLES

Name, address, phone number, SSN, government issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)

Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68AB:6D:35:65:03)

Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses or legal representations)

Clinical or research datasets where only certain research key (e.g., Jane Smith, diabetes, Hgb 15.1 g/dl = C0K123)

Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = S1771561932) (unique sequence not used anywhere else)

Same as Pseudonymous, except data are also protected by safeguards and controls

Data are suppressed, generalized, perturbed, etc. (e.g., GPA, 3.2 + 3.0-3.5, gender: female + gender: male)

Same as De-identified, except data are also protected by safeguards and controls

For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)

Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)

### DIGITAL DATA FLOWS MASTERCLASS: EMERGING TECHNOLOGIES



## City of Seattle Open Data Risk Assessment

JANUARY 2018 – FINAL REPORT



[fpf.org/classes](https://fpf.org/classes)



# Recommended Readings

- **Dep't of Ed/PTAC, Basic Terms Overview -**  
[https://studentprivacy.ed.gov/sites/default/files/resource\\_document/file/data\\_deidentification\\_terms.pdf](https://studentprivacy.ed.gov/sites/default/files/resource_document/file/data_deidentification_terms.pdf)
- **NISTIR 8053: De-Identification of Personal Data** (Oct 2015), <http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf>
- **NIST Special Publication 800-188 (2nd DRAFT): De-Identifying Government Datasets** (Dec 2016),  
[https://csrc.nist.gov/CSRC/media/Publications/sp/800-188/draft/documents/sp800\\_188\\_draft2.pdf](https://csrc.nist.gov/CSRC/media/Publications/sp/800-188/draft/documents/sp800_188_draft2.pdf)
- **UK Anon, Anonymisation Decision-Making Framework -**  
<http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf>
- **Article 29 Working Party (EU guidance body), Opinion on Anonymisation Techniques** (2014),  
[https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216\\_en.pdf](https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf)
- **Ira Rubinstein & Woody Hartzog, Anonymization and Risk**, 91 Wash. L. Rev. 703 (2016),  
<https://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/1589/91WLR0703.pdf>

# UPCOMING WEBINARS



## MODULE 3: USING DATA IN EDUCATION

Friday, May 22nd 3-4 PM ET

## MODULE 4: SHARING DATA

Thursday, June 18th 2-3 PM ET

**Mark Williams**, *Partner, Co-Chair*, eMatters and Higher Education Practice Groups, Fagen Friedman & Fulfrost LLP

# A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.



This is a primer on how to distinguish different categories of data.

## DEGREES OF IDENTIFIABILITY

Information containing direct and indirect identifiers.

## PSEUDONYMOUS DATA


































Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.

## DE-IDENTIFIED DATA

Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.

## ANONYMOUS DATA

Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
<div></div> <b>DIRECT IDENTIFIERS</b> Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)	 INTACT	 PARTIALLY MASKED	 PARTIALLY MASKED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED
<div></div> <b>INDIRECT IDENTIFIERS</b> Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)	 INTACT	 INTACT	 INTACT	 INTACT	 INTACT	 INTACT	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED	 ELIMINATED or TRANSFORMED
<div></div> <b>SAFEGUARDS and CONTROLS</b> Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals	 NOT RELEVANT due to nature of data	 LIMITED or NONE IN PLACE	 CONTROLS IN PLACE	 CONTROLS IN PLACE	 LIMITED or NONE IN PLACE	 CONTROLS IN PLACE	 LIMITED or NONE IN PLACE	 CONTROLS IN PLACE	 NOT RELEVANT due to nature of data	 NOT RELEVANT due to high degree of data aggregation
<b>SELECTED EXAMPLES</b>	Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)	Same as De-Identified, except data are also protected by safeguards and controls	For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)	Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)

# SLDS Technical Brief

*Guidance for Statewide Longitudinal Data Systems (SLDS)*

*December 2010, Brief 3*

*NCES 2011-603*

## Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting

### Contents

Introduction .....	1
Background.....	3
Unintended Disclosure of Personally Identifiable Information .....	4
Current Disclosure Prevention Practices that Retain Some Disclosure Risk.....	7
Best Practices: Practices that Mitigate Disclosure Risk.....	14
Recommendations.....	27
Summary.....	30
References.....	30

*SLDS Technical Briefs are intended to provide “best practices” for consideration by states developing Statewide Longitudinal Data Systems.*

*Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.*

*For more information, contact:  
Marilyn Seastrom  
National Center for Education  
Statistics  
(202) 502-7303  
Marilyn.Seastrom@ed.gov*

### Introduction

Over the last decade, increased attention on education has led to an expansion in the amount of information on students and their schools and school districts reported to parents and the general public (20 U.S.C. § 6311). States now report student outcomes based on assessments of student achievement in specific subjects and grade levels for all students, as well as for subgroups defined by gender, race and ethnicity, English proficiency status, migrant status, disability status, and economic status. Typically, the data are reported as the percentage distribution of students in a subgroup across achievement levels. These reports are issued at the state, district, and school levels. Additional outcome measures, such as data on attendance, dropout rates, and graduation rates, are also reported frequently.

These reports offer the challenge of meeting the reporting requirements while also meeting legal requirements to protect each student’s personally identifiable information (Family Educational Rights and Privacy Act [FERPA]) (20 U.S.C. § 1232g; 34 CFR Part 99). Recognizing this, the reporting requirements state that subgroup disaggregations of the data may not be published if the results would yield personally identifiable information about an individual student (or if the number of students in a category is insufficient to yield statistically reliable information). States are required to define a minimum number of students in a reporting group or subgroup required to publish results consistent with the protection of personally identifiable information (34 CFR § 200.7).

Individual states have adopted minimum group size reporting rules, with the minimum number of students ranging from 5 to 30 and a modal category of 10 (used by 39 states in the most recent results available on state websites in late winter of 2010). Each state has adopted additional practices to protect personally identifiable information about its students in reported results. These practices include various forms of suppression, top and bottom coding of values at the ends of a distribution, and limiting the amount of detail reported for the underlying counts. This Technical Brief includes a summary of key definitions, a brief discussion of background information, and a review and analysis of current practices to illustrate that some practices work better than others in protecting personally identifiable information reported from student education records.

The review led to the formulation of recommended reporting rules that are driven by the size of the reporting groups or subgroups. The reporting rules are intended to maximize the amount of detail that can be safely reported without allowing disclosures from student outcome measure categories that are based on small numbers of students. NCES welcomes input on these recommendations.



## Definitions

**Personally identifiable information** includes the name and address of the student and the student's family; a personal identifier, such as the student's Social Security Number, student number, or biometric record; other indirect information, such as the student's date and place of birth and mother's maiden name; 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 relevant circumstances, to identify a student with reasonable certainty; and information based on a targeted request.

**Disclosure** means to permit access to or the release, transfer, or other communication of personally identifiable information contained in education records by any means. To avoid disclosures in published tables, whenever possible, data about individual students should be combined with data from a sufficient number of other students to disguise the attributes of a single student. When this is not possible, data about small numbers of students should not be published.

**Suppression** refers to withholding information from publication. Some information is withheld from publication in a table to protect data based on small counts because the release of the information would likely lead to a disclosure. Other information is withheld from publication in a table to prevent the calculation of the data based on small counts from the published information; this is known as complementary suppression.

**Recoding** refers to reporting values as being within a specified range rather than as a specific value.

**Top coding** refers to reporting values over a set value as greater than that value.

**Bottom coding** refers to reporting values under a set value as less than that value.

Top coding and bottom coding are specific types of recoding. These procedures are used to protect data for individual students from disclosure.

**Subgroups** refer to students within a larger group who share specific characteristics, such as the subgroup of male students and the subgroup of female students within a school or within a grade in a school. Information from student records is often reported for subgroups of students by gender, race and ethnicity, English proficiency status, migrant status, disability status, and economic status.

**Outcome** measures refer to the student's educational experiences that are recorded in student's educational records. For example, student grades, courses completed, scores on standardized assessments, school attendance, graduation status, participation in extracurricular activities, and disciplinary actions are commonly reported measures of student outcomes.

**Categories** refer to groups of students that share specific experiences that comprise the range of possible outcomes for each educational measure. For example, the percent of students with passing as compared to failing grades, the percent of students who dropout as compared to completing high school, or the percent of students who scored at each of several achievement levels on a standardized state assessment.

## Background

As the nation has focused its attention on education over the last decade, there has been a large increase in the amount of data reported to the general public on America's students and their schools and school districts (20 U.S.C. § 6311(h); 20 U.S.C. § 9607; U.S. Public Law 110-69; U.S. Public Law 111-5). Reporting requirements for public elementary and secondary institutions that receive federal funds include annual status and progress reports at the school, district, and state levels (20 U.S.C. § 6311(h)).<sup>1</sup> Among other requirements, these reports, identified as report cards, must include results from state assessments on the percent of students assessed, along with student achievement results across achievement levels in specific subjects and grade levels for all students and for reporting subgroups including gender, race/ethnicity, English proficiency status, migrant status, disability status, and economic status. The annual status and progress report cards also typically include data on attendance rates and report graduation rates for secondary schools. Dropout rates are also frequently reported at the district and school levels.

The current reporting requirements are typically met through state-, district-, and school-level reports that are published by each state's department of education. These reports offer the challenge of balancing the reporting requirements against legal requirements to protect each student's personally identifiable information (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99). To this end, the reporting requirements for Title I state that disaggregating the data for specific subgroups may not occur if the number of students in a reporting group or subgroup is insufficient to yield statistically reliable information or if the results would yield personally identifiable information about an individual student (20 U.S.C. § 6311(h); 34 CFR § 200.7).<sup>2</sup>

As part of the reporting requirements, each state is required to have an accountability plan that describes its system for monitoring adequate yearly progress with annual objectives for continuous and substantial improvement for all students and for each specified student subgroup. In addition to defining specific measures, each state's accountability plan is expected to include the state's definition of the minimum number of students in a subgroup required for reporting purposes and information as to how the State Accountability System protects the privacy of students when reporting results.

What does protecting student privacy mean in a reporting context? In order to protect a student's privacy, the student's personally identifiable information must be protected from public release. The broad, federal government-wide definition of personally identifiable information states "the term 'personally identifiable information' refers to information that can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc., alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc." (OMB Memorandum 07-16, *Safeguarding Against and Responding to the Breach of Personally Identifiable Information; Implementation Guidance for Title V of the E Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)*).

---

<sup>1</sup> The requirement specified in law is for an annual state report card and for annual district report cards that include information for the district and each school.

<sup>2</sup> The law states that reporting student assessment results disaggregated by economically disadvantaged students, students from major racial and ethnic groups, students with disabilities, and students with limited English proficiency is not required if the number of students in a *category* is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student (20 U.S.C. § 6311). However, the regulations use the term *subgroup* to refer to the disaggregated student data, and the regulations specify that a state may not report achievement results for a *subgroup* if the results would reveal personally identifiable information about an individual student (34 CFR § 200.7). This is further promulgated in the September 12, 2003 non-regulatory guidance on Report Cards Title I, Part A.



The FERPA definition of personally identifiable information (34 CFR § 99.3) follows the

government-wide definition and includes the following:

*Personally identifiable information includes, but is not limited to:*

1. The student's name;
2. The name of the student's parent or other family members;
3. The address of the student or student's family;
4. A personal identifier, such as the student's Social Security Number, student number, or biometric record;<sup>3</sup>
5. Other indirect identifiers, such as the student's date of birth, place of birth, and mother's maiden name;
6. 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;
7. Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates.  
(34 CFR § 99.3)

Protecting student privacy means publishing data only in a manner that does not reveal individual students' personally identifiable information, either directly or in combination with other available information. Another way of putting this is that the goal is to publish summary results that do not allow someone to learn information about a specific student.

States publish annual status and progress reports that are based on reports of outcome measures at the school, district, or state level. These reports aggregate, or combine, the results for individual students into summary statistics. These statistics include the number or percentage of students overall or in each of the reporting subgroups for specific outcome measures (e.g., the percentage of students in each racial and ethnic group who graduate from high school; the percentage of English language learners who score in each achievement level on a state assessment).

This report demonstrates how disclosures occur even in summary statistics. It describes

various reporting practices and data protection techniques currently in use and illustrates how commonly used methods of data protection may fall short of their goal. The report then identifies "best practices" to avoid the unintended disclosure of personally identifiable information, including publishing the percentage distribution across categories of outcome measures with no underlying counts or totals; publishing a collapsed percentage distribution across categories of outcome measures with no underlying counts or totals; publishing counts but using complementary suppression at the subgroup level when a small subgroup is suppressed; limiting the amount of detail published for school background information; recoding the ends of percentage distributions; and recoding high and low rates. This information is used to develop recommendations for reporting rules that maximize the amount of information reported while protecting the privacy of each student's data.

## Unintended Disclosure of Personally Identifiable Information

When personally identifiable information is revealed through information released to the public, it is called a disclosure.<sup>4</sup> When

schools, districts, or states release information about educational progress, they typically release aggregated data—data for groups of

<sup>3</sup> FERPA 2008 regulations state that the term "biometric record, as used in the definition of personally identifiable information, means a record of one or more measurable biologic or behavioral characteristics that can be used for automated recognition of an individual. Examples include fingerprints; retina and iris patterns; voiceprints; DNA sequence; facial characteristics; and handwriting." (34 CFR § 99.3)

students—to prevent disclosure of information about an individual. Even with some methods of aggregation, unintended disclosure of personally identifiable information may occur. How could data reporting outcome measures for groups of students possibly reveal information on an individual student? The example that follows shows how information about individual students' achievement levels can be revealed, even in data reported for groups of students. Furthermore, it shows that the identity of groups of students can be revealed within combinations of achievement levels (e.g., Below Basic and Basic for students who scored below Proficient, or proficient and advanced for students who scored at or above Proficient).

Typically, each child's parents are given their child's score and achievement level on the state assessment as well as the report for their child's school. Table 1 provides the percentage distribution and number of students at each achievement level at the school level in grade 4 mathematics, for students overall and for several subgroups: White and Hispanic students, students with and without an individualized education plan, and students who are and are not English language learners. Any combination of these three subgroup variables that reveals the achievement level for a student or group of students with identifiable characteristics results in a disclosure.

### ***Example 1: Unintended Disclosures***

Consider a school report that includes results on the state assessment by grade and subject. No results are suppressed as a result of a small subgroup count, since each subgroup included more than the minimum reporting group size of 5. The report shows that there are 32 fourth-graders in this school and that they were all assessed in mathematics (table 1). Among these students, 12.5 percent, or 4 students, scored at the Below Basic achievement level; 31.3 percent, or 10 students, scored at the Basic level; 34.4 percent, or 11 students, scored at the Proficient level; and 21.9 percent, or 7 students, scored at the Advanced level. The data reported for the subgroups of students with and without an individualized education plan show that all fourth-graders with an individualized education plan scored below the Proficient level (4 students at the Below Basic level plus 3 at the Basic level). Assuming that other students in the class know who among their peers have individualized education plans, this is a disclosure because it reveals that each fourth-grader with an individualized

education plan failed to reach the Proficient level on the assessment.

Next, looking at the 10 Hispanic fourth-graders, the data show that 1 student in this subgroup scored at the Proficient level, while the other 9 students scored at either the Basic level (5 students) or the Below Basic level (4 students). Since parents receive their child's score and achievement level as well as a school report that shows the performance in mathematics by grade, the parents of the 1 Hispanic student who scored at the Proficient level know that the other 9 Hispanic students in the fourth grade each scored below the Proficient level in mathematics. This is a disclosure, because these parents now know that each of their child's ethnic peers failed to reach the Proficient level.<sup>5</sup>

The subgroup data in this table also show that each of the 4 fourth-graders who scored at the Below Basic level were Hispanic, received English language instruction, and had an individualized education plan. This is a considerable amount of information

about the characteristics of the 4 lowest performers. However, since there were Hispanic students who scored at the Below Basic, Basic, and Proficient achievement levels, students with individualized education plans who scored at both the Below Basic and Basic achievement levels, and students receiving English language instruction who scored at both the Below Basic and Basic achievement levels, the table only identifies the fact that there are four Hispanic fourth-graders with this set of three shared characteristics; it does not identify the 4 specific Hispanic students. Thus, the table considered alone does not result in a disclosure in this instance.

Suppose, however, that the students with individualized education plans receive observable special services (e.g., a tutor, extra time on tests, one-on-one test instruction) and that there are exactly 4 Hispanic students receiving these services; then it becomes apparent that these are the 4 Hispanic students who scored at the Below Basic achievement level.

<sup>4</sup> Under FERPA, disclosure means to permit access to or the release, transfer, or other communication of personally identifiable information contained in education records by any means, including oral, written, or electronic means, to any party except the party identified as the party that provided or created the record (34 CFR § 99.3).

<sup>5</sup> While this disclosure is based on the parents' personal knowledge of their child's score, the fact that each parent in the school receives his or her child's score raises this source of disclosure as a topic of concern (i.e., knowledge of one child's score revealing the performance of other students).

**Table 1. School-level grade 4 mathematics assessment results in a state with a minimum reporting group size of 5**

		Percent assessed	Tested	Below Basic	Basic	Proficient	Advanced
<b>Total</b>	%	100	100	12.5	31.3	34.4	21.9
	N	†	32	4	10	11	7
White	%	100	100	0.0	22.7	45.5	31.8
	N	†	22	0	5	10	7
Hispanic	%	100	100	40.0	50.0	10.0	0.0
	N	†	10	4	5	1	0
Individualized education plan	%	100	100	57.1	42.9	0.0	0.0
	N	†	7	4	3	0	0
No individualized education plan	%	100	100	0.0	28.0	44.0	28.0
	N	†	25	0	7	11	7
English language learner	%	100	100	40.0	50.0	10.0	0.0
	N	†	10	4	5	1	0
Not English language learner	%	100	100	0.0	22.7	45.5	31.8
	N	†	22	0	5	10	7

† Not applicable.

NOTE: Details may not sum to totals because of rounding.

Recall that the reporting requirements acknowledge the risk associated with small numbers by indicating that results should only be published if the results would not reveal personally identifiable information about an individual student. The instructions for the state

accountability plan also acknowledge this risk with the requirement for each state to establish a minimum subgroup size for reporting and with the requirement for each state to describe how the State Accountability System protects the privacy of students when reporting results.

## Current Disclosure Prevention Practices that Retain Some Disclosure Risk

Typically, a state establishes the required minimum number of students in a subgroup for privacy protection and then does not report the results for outcome measures for any subgroup with less than this established minimum number. The groups not reported are identified as having been suppressed to protect student privacy. A review in late winter of 2010 of the most recent reported assessment results for each state and the District of Columbia found that 39 states use a minimum reporting group size of 10 students. Another 7 states set the minimum reporting group size at 5, and 5 states set the minimum higher, with values ranging from 15 to 30.

While subgroup suppression is a good start, it may not be enough to prevent disclosure of personally identifiable information. The descriptions of current practices include such potentially problematic methods as 1) suppressing data for small subgroups but not for small categories of outcome measures for reported subgroups; 2) suppressing data for small subgroups but reporting counts across the categories of the outcome measure for the overall group and the reported subgroups; 3) suppressing data for small subgroups but reporting the overall total count; and 4) suppressing data for small subgroups but reporting ranges for the overall totals and the reported subgroup totals.

### Suppressing Data for Subgroups but not for Reporting Categories

The practice of suppressing data for small subgroups is a start. However, when subgroup results are reported for the categories of an outcome measure, there can also be a small number of students in one or more of the categories within the larger subgroups. Reporting results for small numbers of students within a category or within a subgroup can present a risk to student privacy because it increases the risk of unintentionally releasing information that identifies individual students. The minimum for categories within subgroups can be set lower

than the size of the subgroup minimum, but there should be a minimum size specified for individual categories to guard against unintentional disclosures. This minimum, which is sometimes referred to as the threshold rule, defines those categories in a table that are defined as sensitive because the number of students is less than the specified number. Some data collection agencies set this number at 5, while others set it as 3. (Federal Committee of Statistical Methodology, Working Paper 22). Sensitive categories are illustrated in the following example.

## Example 2: Suppression of Small Subgroups but not Small Categories

In this example, when a minimum reporting size of 10 is applied to the data from table 1, the assessment results for the 7 students with individualized education plans are presumed to be protected from disclosure because the results are suppressed (see table 2). Thus, the result in example 1 showing that all students with an individualized education plan failed to reach the Proficient level of the state assessment is presumed to be protected from

disclosure. However, when the assessment results of the 10 Hispanic students and the 10 English language learners are reported across the four achievement levels, the number of students at each achievement level falls below the established minimum reporting size. In both subgroups, there are 4 students in the Below Basic achievement group, 5 students in the Basic achievement group, and 1 student in the Proficient achievement group; nevertheless, the results are

reported since the minimum size rule is applied at the subgroup reporting level. As described in example 1, reporting that only one Hispanic child scored at or above the Proficient level discloses information about that child and about the achievement level of the other students in the subgroup. Anyone who is able to identify the Hispanic child with a high score then knows that the other Hispanic children in the same grade failed to reach the proficient achievement level.

**Table 2. School-level grade 4 mathematics assessment results in a state with a minimum reporting group size of 10**

		Percent assessed	Tested	Below Basic	Basic	Proficient	Advanced
<b>Total</b>	%	100	100	12.5	31.3	34.4	21.9
	N	†	32	4	10	11	7
White	%	100	100	0.0	22.7	45.5	31.8
	N	†	22	0	5	10	7
Hispanic	%	100	100	40.0	50.0	10.0	0.0
	N	†	10	4	5	1	0
Individualized education plan	%	100	100	*	*	*	*
	N	†	7	*	*	*	*
No individualized education plan	%	100	100	0.0	28.0	44.0	28.0
	N	†	25	0	7	11	7
English language learner	%	100	100	40.0	50.0	10.0	0.0
	N	†	10	4	5	1	0
Not English language learner	%	100	100	0.0	22.7	45.5	31.8
	N	†	22	0	5	10	7

† Not applicable.

\* Not reported to protect subgroups with fewer than 10 students.

NOTE: Details may not sum to totals because of rounding.

## Suppressing Data for Subgroups but Reporting Too Much Detail in Underlying Counts

Suppressing data for small subgroups is a first step. However, when data are suppressed to protect student privacy, care must also be taken to avoid publishing information that can be used to retrieve or recover the suppressed information. The next three examples illustrate disclosure problems that can occur in reporting student outcome measures.

The released data in each example table are displayed with a white background. The shaded portions of the example tables represent data that were suppressed. The data entries in the shaded portions of the table were recovered from the released data.

### *Counts for overall group and reported subgroups*

In 38 states, the data are suppressed for subgroups that fall below the minimum reporting group size; however, the number of students and the percentage distributions across the categories of the outcome measure are reported for the overall group and the remaining reporting subgroups. The reported information can then be used to recover the suppressed data through a series of calculations. This can be done using the following steps:

1. Convert the percentages across the outcome categories for the overall group to proportions.
2. Multiply the proportions by the number of students in the overall group to yield the number of students in each category of the outcome measure in the overall group.
3. Identify a suppressed subgroup and the related reported subgroup(s).
4. Repeat steps 1 and 2 for the related reported subgroup(s) to yield the number of students in each category of the outcome measure in the reported subgroup.
5. Subtract the number of students in each category of the outcome measure for the reported subgroup from the overall count for that outcome category to yield the number of students in each category of the outcome measure for the suppressed subgroup.
6. If there are more than 2 subgroups for one disaggregation (e.g., race/ethnicity), compute

the counts across the categories of the outcome measure for each reported subgroup, sum subgroup counts for the reported subgroups across each outcome category, and then subtract from the overall number for that category of the outcome measure to yield the number of students in each category of the outcome measure for the suppressed subgroup(s).

All students are in one of two subgroups when student outcome measures are reported by gender, economic status, English proficiency status, migrant status, or disability status. When the data for one of the two subgroups are suppressed and the data for the other subgroup and the total are published, the suppressed data can be fully recovered. When student outcome measures are reported for race and ethnicity, subgroup data are frequently suppressed for more than one subgroup. However, the difference between the counts computed for the outcome categories of students overall and the summation across the outcome categories for the reported subgroups can be used to recover data for the suppressed subgroup(s). This recovery may yield identifying information about the students in the reporting subgroup(s) with suppressed data.

The recovery of suppressed results does not always pose a serious threat to students' personally identifiable information, but in some instances it does—the risk of identifying an individual student is a function of the distribution of students across the recovered categories.

### ***Example 3: Suppressing Outcomes but Reporting Counts for Subgroups***

The reported data in table 3 show that among 82 students who were assessed in third-grade reading, 7.3 percent (6 students) scored at the Below Basic achievement level, 42.7 percent (35 students) scored at the Basic level, 37.8 percent (31 students) scored at the Proficient level, and 12.2 percent (10 students) scored at the Advanced level. Seventy-five of the 82 students did not have an individualized education plan, and the reported data show that 8.0 percent (6 students) in this reporting subgroup scored at the Below Basic level, 42.7 percent (32 students) scored at the Basic level, 36.0 percent (27 students) scored at the Proficient level, and 13.3 percent (10 students) scored at the Advanced level.

Although the data were suppressed for students with an individualized education plan, the recovered data show that 7 of the 82 students

assessed in third-grade reading were in this suppressed reporting subgroup. Further, a comparison of the overall assessment results with those for the 75 students without an individualized education plan shows that 3 of the 7 students with an individualized education plan scored at the Basic level and 4 scored at the Proficient level. These data do not provide the information needed to identify which students with an individualized education plan scored at the Proficient level and which did not. Thus, this table does not disclose an individual student's performance; however it does reveal the fact that no student with an individualized education plan scored at the Advanced level or at the Below Basic level.

In contrast, the recovered data for 8 low-income students show that 3 of these students scored at the Below Basic achievement level and 5 scored

at the Basic achievement level. Thus, all students identified as low-income scored below the Proficient achievement level. If an individual student is known to be from a low-income family, the information in this table discloses that student's score as below Proficient.

The recovered data for 8 students receiving English language instruction show that 3 scored at the Below Basic achievement level, 4 scored at the Basic achievement level, and 1 scored at the Proficient level. Since parents receive their child's score along with the school report, the parents of the child who scored at the Proficient level could use the information in the published table for their child's grade to learn that each of their child's peers who received English language instruction failed to score at the Proficient achievement level.

**Table 3. School-level grade 3 reading assessment results for a state with a minimum reporting size of 10**

		Tested	Below Basic	Basic	Proficient	Advanced
<b>Total</b>	%	100	7.3	42.7	37.8	12.2
	N	82	6	35	31	10
Individualized education plan	%	100	0.0	42.9	57.1	0.0
	N	7	0	3	4	0
No individualized education plan	%	100	8.0	42.7	36.0	13.3
	N	75	6	32	27	10
English language learner	%	100	37.5	50.0	12.5	0.0
	N	8	3	4	1	0
Not English language learner	%	100	4.1	41.9	40.5	13.5
	N	74	3	31	30	10
Low income	%	100	37.5	62.5	0.0	0.0
	N	8	3	5	0	0
Not low income	%	100	4.1	40.5	41.9	13.5
	N	74	3	30	31	10

NOTE: Details may not sum to totals because of rounding.



### Counts for the overall group

Some states report the percentage distribution across achievement levels for the overall population in a grade and subject along with the percentage distributions for each subgroup, but only publish the number of students tested overall for that grade and subject. This seems like it would provide more protection to students' personally identifiable information, since the number of

students in each subgroup is not published. However, in many cases—especially at the school or district level for the data reported by grade and subject—there is only one unique mathematical solution that could yield the reported subgroup percentage distributions for the reported number of students overall.

### Example 4: Suppressing Outcomes but Reporting Counts for Groups

In this school, 46 students were assessed in third-grade reading (table 4), and this number is known. Note that the shaded cells in the table display the data that were recovered from the reported information. Multiplying the proportions from the percentage distribution times the number in the overall group (46) shows that the 6.5 percent who scored at the Below Basic level represents 3 students (i.e.,  $0.65 \times 46 = 3$ ). The data reported by gender show that the 3 students who scored at the Below Basic level are all males. Thus, by dividing 8.3 by 3, the data show that each male student represents 2.77 percent of the number of males in the subgroup. Dividing each of the

remaining percentages by 2.77 shows that there are 10 males who scored at the Basic level, 20 who scored at the Proficient level, and 3 who scored at the Advanced level.

Next, the number of males at each achievement level is subtracted from the number of students at that achievement level to recover the suppressed data for females. These calculations show that there are no females at the Below Basic level, no females at the Basic level, 7 females at the Proficient level, and 3 females at the Advanced level. The recovered data do not reveal which females scored at each of these two levels. However, when the focus of the

reporting or interpretation of the data shifts to performance at or above versus below the Proficient level, the data for students scoring at the Below Basic and Basic level are combined to show the percent of students who scored below the Proficient level and the percent of students who score at the Proficient and Advanced levels are combined to show the percent of students who scored at the Proficient level. In this example, the recovered data show that all of the third-grade females in this school scored at the Proficient level or above in reading. This then discloses information about the reading achievement level of each of the third-grade females in this school.

**Table 4. School-level grade 3 reading assessment results for a state with a minimum reporting size of 10**

		Tested	Below Basic	Basic	Proficient	Advanced
Total	%	100	6.5	21.7	58.7	13.0
	N	46	3	10	27	6
Male	%	100	8.3	27.8	55.6	8.3
	N	36	3	10	20	3
Female	%	100	0.0	0.0	70.0	30.0
	N	10	0	0	7	3

NOTE: Details may not sum to totals because of rounding.

## Counts for the overall group and subgroups reported as ranges

Another reporting approach recognizes the problem with reporting exact population counts for students assessed and, instead, reports the counts in ranges (i.e., as a categorical variable). With this approach, the percentage distribution is reported for each grade and subject overall and for each of the reporting subgroups that do not require suppression; then, instead of reporting the exact number of students in each group or subgroup, a range that includes the exact number is all that is reported for the count (e.g., instead of reporting 33 students, the number is reported

as 30–39). As with the last approach, this would seem to provide more protection to students' personally identifiable information, since the exact number of students is not published. However, the range of possible values for the number of students can be used to identify the number of students that, when applied to the proportion of students at each achievement level, yields estimates that are the closest to whole numbers. Once these counts are established for the overall group and for a reported subgroup, the suppressed counts for a related subgroup can be recovered.

### Example 5: Suppressing Outcomes but Reporting Ranges for Counts

The number of third-graders assessed in reading was reported as 40–49 (table 5). The percentage distribution of third-graders overall, across the achievement levels, was reported with 2 decimal places. The percentage distribution across the achievement levels was reported for the 30–39 students who did not have an individualized education plan, but the achievement results were suppressed for the 6–9 students who had one. First, the proportions from the distribution across the achievement levels were applied to each of the 10 numbers in the 40 to 49 range. The number that resulted in estimates that were closest to whole numbers is 41. This showed that, overall, 2

students scored at the Below Basic level, 5 scored at the Basic level, 15 scored at the Proficient level, and 19 scored at the Advanced level. Next, this set of steps was repeated for the 10 numbers in the 30–39 range, using the proportions from the percentage distribution across the achievement levels for students who did not have an individualized education plan. This showed that there were 34 students in this group, with none at the Below Basic level, none at the Basic level, 15 at the Proficient level, and 19 at the Advanced level.

Finally, the counts for students who did not have an individualized education plan were subtracted from

the overall counts to recover the suppressed number for the students with an individualized education plan—there were 7 students in this group. Within this group, 2 scored at the Below Basic level, 5 scored at the Basic level, none scored at the Proficient level, and none scored at the Advanced level. These counts can then be used to compute the suppressed percentage distribution. The recovered data show that each of the 7 third-graders with individualized education plans scored below the Proficient level in reading. This is a disclosure of the reading achievement-level information for these 7 students

**Table 5. School-level grade 3 reading assessment results for a state with a minimum reporting size of 10 and counts reported as ranges**

		Percent assessed	Number tested	Below Basic	Basic	Proficient	Advanced
<b>Total</b>	%	100	†	4.88	12.20	36.59	46.34
	N	40–49	41	2	5	15	19
Individualized education plan	%	100	†	28.57	71.43	0.00	0.00
	N	6–9	7	2	5	0	0
No individualized education plan	%	100	†	0.00	0.00	44.12	55.88
	N	30–39	34	0	0	15	19

† Not applicable.

NOTE: Details may not sum to totals because of rounding.

## Best Practices: Practices that Mitigate Disclosure Risk

The review of each state's online reporting of assessment results for schools uncovered three approaches that can help in protecting against the release of information needed to recover personally identifiable information. The first such approach involves not reporting any of the enrollment data that were used to compute the percentage distributions across the achievement-level results. The second approach starts with the first approach (i.e., the underlying enrollment counts are not reported) and collapses across outcome categories to further limit the amount of detail published. This increases the number of students included in each reported outcome category. The third approach involves suppressing subgroups other than the subgroups with less than the minimum reporting size in order to prevent the recovery of the suppressed results for the small subgroups.

### No Counts Published

Eight states were identified that publish student assessment results by grade and subject for the overall student population and for the reportable subgroups (i.e., those subgroups that do not require suppression) only as a percentage distribution across the achievement levels. In these states, the school reports do not include counts of the number of students assessed overall or of the number of students assessed in each of the reporting subgroups. However, since too much precision in the percentages can limit the possible options for the underlying counts, limiting the

Additional practices that support public reporting while protecting student privacy were identified and are discussed in this section. The first involves the reporting of background data on enrollment by grade and enrollment by student characteristics for a school or district. The second involves protecting data at the ends of the distribution, or at the low and high values for a rate, to avoid reporting that a small number of students (or nearly all students) have a specific outcome.

Each of these practices taken alone does not necessarily address each of the potential sources of disclosure, but they do reflect practices that, when taken in combination, may lead to improved protection of personally identifiable information about individual students in published tables.

percentages reported to whole numbers increases the number of possible options for the underlying counts. This helps protect the suppressed data for small groups. It also helps protect the counts for small categories within outcome measures for the reported subgroups. The following example of school-level third-grade reading results shows that while the relative relationships across achievement levels within and across subgroups are evident, the absence of the counts used to compute the percentage distributions prevents the recovery of the suppressed data.

### Example 6: Best Practices: No Counts Published

Table 6 shows assessment results only as percentage distributions reported as whole numbers. This, coupled with the fact that no counts are reported, protects the suppressed data from disclosure (table 6). The table shows that 13 percent of the students scored

at the Below Basic level, 44 percent scored at the Basic level, 27 percent scored at the Proficient level, and 16 percent scored at the Advanced level. Relatively more male than female students and more low-socioeconomic status than non-low-socioeconomic

status students performed at the Below Basic level. The data are suppressed for the English language learner subgroup because there are fewer than 10 students in the subgroup.

**Table 6. Percentage distribution of school-level grade 3 reading assessment results in a state with a minimum reporting size of 10 and no counts**

	Below Basic	Basic	Proficient	Advanced
<b>Total</b>	13	44	27	16
Male	17	47	23	13
Female	9	42	30	18
Low SES	28	39	22	11
Not low SES	7	47	29	18
English language learner	*	*	*	*
Not English language learner	6	44	31	19

\* Not reported to protect subgroups with fewer than 10 students.

NOTE: Details may not sum to totals because of rounding. SES = Socioeconomic status.

**Collapsing Across Outcome Categories**

Seven states limited their reporting of achievement results to two categories—those at or above the level established by the state for successful performance and those who did not score in the successful range. Collapsing across outcome categories is useful when there are a small number

of students in one or more of the outcome categories. This approach, combined with the decision to not report the underlying counts, is another way of increasing the protection of student privacy in reported summary tables.

**Example 7: Best Practices: Collapsing across Outcome Categories**

Collapsing across outcome categories and displaying the assessment results only as a percentage distribution protects the underlying counts from disclosure. Collapsing the data used in the previous example, 57 percent

of the students scored at or below the Basic level, and 43 percent scored at or above the Proficient level (table 7). Relatively more male than female students (64 percent versus 51 percent) and low socioeconomic status than

not low socioeconomic status students (67 percent versus 53 percent) scored at the Below Basic level. The data are suppressed for the English language learner subgroup because there are less than 10 students in the subgroup.

**Table 7. Percentage distribution of school level, grade 3 reading assessment results collapsed in a state with a minimum reporting size of 10 and no counts**

	Basic or below	Proficient or above
Total	57	43
Male	64	36
Female	51	48
Low SES	67	33
Not low SES	53	47
English language learner	*	*
Not English language learner	50	50

\* Not reported to protect subgroups with fewer than 10 students.  
NOTE: Details may not sum to totals because of rounding. SES = Socioeconomic status.

## Counts Published with Additional Suppression

One state provides counts for the overall number of students assessed in a specific grade and subject and for students in reportable subgroups. However, instead of suppressing only the subgroups that do not meet the minimum reporting size, subgroups related to the suppressed group are also suppressed. This is referred to as “complementary suppression.” That is, a subgroup

with less than 10 students is suppressed, and one (or more) of the other subgroups that combine with the small subgroup to account for a larger share of the students in the overall group is also suppressed. The following example of school-level third-grade reading results provides an illustration of this approach.

### ***Example 8: Best Practices: Schools Counts Published with Additional Suppression***

This example includes two schools. The school-level report is designed to display results by gender, race and ethnicity, low-income status, and individualized education plan status. School 1, with 30 students, had a number of reporting subgroups with fewer than 10 students. Suppressing the assessment results for the small subgroups and suppressing the outcome measure for a related category (i.e., complementary

suppression of additional rows of the table) protects the reported data at the school level, but leads to the loss of information. As shown in table 8, data were suppressed for the 27 White students because there were fewer than 10 students in each of the other racial and ethnic subgroups (i.e., 2 Native American students and 1 Black student). Data were suppressed for the 21 low income students because there were fewer than 10 students

who were not low income. Data were also suppressed for the 21 students without an individualized education plan, because only 9 students had individualized education plans. By comparison, assessment data were reported for the 30 third-grade students overall, and for the 12 male and 18 female students because the minimum reporting threshold of 10 students was exceeded in each case.

**Table 8. School 1: Number tested and percentage distribution of grade 3 reading assessment results with a minimum reporting size of 10 and complementary row suppression**

	Number tested	Below Basic	Basic	Proficient	Advanced
<b>Total</b>	30	16.7	56.7	20.0	6.7
Male	12	25.0	58.3	16.7	0.0
Female	18	11.1	55.6	22.2	11.1
White	27	*	*	*	*
Native American	2	*	*	*	*
Black	1	*	*	*	*
Low income	21	*	*	*	*
Not low income	9	*	*	*	*
Individualized education plan	9	*	*	*	*
No individualized education plan	21	*	*	*	*

\* Not reported to protect subgroups with fewer than 10 students.

School 2, with 45 students, had 10 or more students in each reporting group. As a result, no data were suppressed

and the third-grade reading assessment results were reported for each of the reporting variables—gender, race

and ethnicity, low income status, and individualized education plan status (table 9).

**Table 9. School 2: Number tested and percentage distribution of grade 3 reading assessment results with a minimum reporting size of 10 and complementary row suppression**

	Number tested	Below Basic	Basic	Proficient	Advanced
<b>Total</b>	45	2.2	22.2	62.2	13.3
Male	18	5.6	27.8	55.6	11.1
Female	27	0.0	18.5	66.7	14.8
White	20	0.0	10.0	65.0	25.0
Native American	10	10.0	40.0	50.0	0.0
Black	15	0.0	26.7	66.7	6.7
Low income	14	7.1	21.4	64.3	7.1
Not low income	31	0.0	22.6	61.3	16.1
Individualized education plan	11	9.1	72.7	18.2	0.0
No individualized education plan	34	0.0	5.9	76.5	17.6

These two schools are the only schools in a district that include the third grade. When the data for the two schools were combined at the district level,

there were 10 or more students in each reporting group. The resulting data are displayed in the next example.

**Example 9: Best Practices:  
District Counts Published  
with Additional Suppression**

Since there were more than 10 students in each reporting subgroup at the district level, the district table based on the schools in example 8 (tables 8 and 9) was produced with full details reported for each reporting group. Table 10 displays these results.

**Table 10. Number tested and percentage distribution of district-level grade 3 reading assessment results with a minimum reporting size of 10 and complementary row suppression**

	Number tested	Below Basic	Basic	Proficient	Advanced
<b>Total</b>	75	8.0	36.0	45.3	10.6
Male	30	13.4	40.0	40.0	6.7
Female	45	4.4	33.3	48.9	13.3
White	47	6.4	38.3	40.4	14.9
Native American	12	16.7	41.7	41.7	0.0
Black	16	6.3	25.9	62.5	6.3
Low income	35	17.1	54.3	25.7	2.8
Not low income	40	0.0	20.0	62.5	17.5
Individualized education plan	20	30.0	55.0	15.0	0.0
No individualized education plan	55	0.0	29.1	56.4	14.5



But with all of the details published for school 2 and for the district, the percentage distribution across the achievement levels in each row can be converted to proportions. The proportions can then be applied to the number of students in the reporting subgroup to compute the number of students at each achievement level in each reporting group. Once this is done at the district level and for school 2, all of the suppressed data for school 1 can be recovered. For example, 38.3 percent of the 47 White third graders in the district scored at the Basic achievement level. Multiplying 0.383 times 47 shows that 18 White third graders in the district scored at the Basic achievement level. The results for White third graders in school 2 show that 10 percent of the 20

students in this subgroup scored at the Basic achievement level. Multiplying 0.10 times 20 shows that 2 White third graders in School 2 scored at the Basic achievement level. Subtracting the 2 students from School 2 from the 18 students in the district reveals the fact that there were 16 White third graders in School 1 who scored at the Basic achievement level. These 16 students comprise 59.3 percent of the 27 White third graders in school 1. These procedures were repeated to recover each of the percentages that were suppressed for school 1 in table 8. The recovered results for school 1 are shown in the shaded cells in table 11 which show that the 2 Native American third graders scored at or below Basic, the 1 Black third grader scored below Basic, and 23.8

percent of the 21 low income students scored below Basic and the other 76.2 percent scored at the Basic level. When the results for students who scored at the below Basic and Basic levels are combined to show the percent who scored below proficient, the data show disclosures of the fact that all students who were Native American, Black, or low income scored below the Proficient level. Furthermore, the parents of the 1 third grade student in school 1 with an individualized education plan who scored at the Proficient achievement level (i.e., 11.1 percent of 9 students is 1 student) know that the other third graders with individualized education plans each failed to reach the Proficient achievement level.

**Table 11. School 1: Number tested and percentage distribution of grade 3 reading assessment results with suppressed percents recovered**

	Number tested	Below Basic	Basic	Proficient	Advanced
<b>Total</b>	30	16.7	56.7	20.0	6.7
Male	12	25.0	58.3	16.7	0.0
Female	18	11.1	55.6	22.2	11.1
White	27	11.1	59.3	22.2	7.4
Native American	2	50.0	50.0	0.0	0.0
Black	1	100.0	0.0	0.0	0.0
Low income	21	23.8	76.2	0.0	0.0
Not low income	9	0.0	11.1	66.7	22.2
Individualized education plan	9	55.6	33.3	11.1	0.0
No individualized education plan	21	0.0	66.7	23.8	9.5

\* Not reported to protect subgroups with fewer than 10 students.

This example illustrates the fact that it is not enough to simply suppress results at the school level, since comparisons of data published for other schools and the district can be used to recover suppressed results within a school. To avoid the recovery of suppressed school level results, the results for other schools in the district and the results for the district must also be taken into account. If the results for a specific subgroup are suppressed in at least two schools, the suppressed results for each school cannot be recovered from the results reported for other

schools and the district. However, when the results are suppressed for a specific subgroup in only one school, to protect the suppressed results from recovery, the results for that subgroup must be suppressed for either another school in the district or for the district.

To protect results that are suppressed at the district level, the same precautions must be taken across district and state results. To protect suppressed results from recovery, if the results are suppressed for a specific subgroup in one district, the results

for that subgroup must be suppressed for a second district in the state.

It is important to note that this problem is not limited to applications that use complementary suppression across related subgroups. The same comparisons between district results and the results reported for other schools in the district or between state results and the results reported for other districts in the state can be applied when the results are suppressed for a single subgroup (i.e., without complementary subgroup suppression).

Care must be taken to ensure that the suppressed results for a subgroup in a single school or single district cannot be recovered using reported data for other schools in the district or other districts in the state. This can be achieved by ensuring that the results for a suppressed subgroup are suppressed in two schools. Alternatively, in districts with only one school for a grade, the results for the suppressed subgroup must also be suppressed at the district level. Similarly, the results for a suppressed subgroup must be suppressed for two districts in a state.

**Reporting School-, District-, or State-Level Background Information**

In reports of outcome measures, some school-, district-, or state-level reports display background information on the distribution of students in a school, district, or state in two separate summary tables. One summary table reports the total number of students enrolled and the percentage of students enrolled by grade. The second summary table reports the total number of students enrolled and the percentage of students in each of the reporting subgroups (e.g., gender, race and ethnicity, English proficiency status, migrant status, disability status, and economic status). Thus, rather than providing the exact number or percentage of students in each grade in each reporting subgroup, the report gives a portrait of the school, district, or state. However, if the number of students reported for an individual grade is the same as the number of students enrolled on the assessment date, that number, along with the report of the percentage of the students who participated in the assessment, can

be used with the percentage distribution across the achievement levels to recover the underlying numbers of students who scored at each achievement level.

Three things can be done to counter this problem. First, use background enrollment counts for a day other than that of the assessment administration and clearly label the date of the background enrollment counts and the date of the assessment in public reports to establish the fact that they are different. Second, report the percentage distribution for the background data and for the results reported across the achievement levels only in whole numbers. This decreases the precision of the reported percentages, which lowers the chance of an accurate recovery of the numbers of students in both reported and suppressed results. Third, report the percentage of students assessed as a whole number.

**Example 10: Best Practices: Reporting Background Information**

Table 12 provides an example of school-level data for enrollment by grade for an elementary school with grades K–6. The shaded cells are not included in the reported table, but are included here to illustrate the added protection from reporting the percentage distribution without any decimal places. For example, 4 of the 7 grades are reported as being 14

percent of the school’s enrollment; the underlying data show that the more precise percentages are 13.9, 14.5, 13.6, and 14.2. The state assessment in this state is administered in March of each school year; reporting enrollment data from 5 months earlier in the school year is likely to result in some differences from the enrollment data at the time of the assessment.

Table 13 displays school-level enrollment data reported by student characteristics for the same elementary school. Again, the patterned cells are not included in the reported table. Taken together, these tables provide a profile of the school without providing the level of detail needed to recover the underlying counts for the outcome measures reported for the school.

**Table 12. Elementary school enrollment, by grade**

	Number	Unrounded percent	Percent
<b>Total</b>	359	100.0	100
Kindergarten	50	13.9	14
Grade 1	52	14.5	14
Grade 2	54	15.0	15
Grade 3	49	13.6	14
Grade 4	48	13.4	13
Grade 5	51	14.2	14
Grade 6	55	15.3	15

**Table 13. Elementary school enrollment, by selected characteristics**

	Number	Unrounded percent	Percent
<b>Total</b>	359	†	†
Male	185	51.5	52
Female	174	48.5	48
White	221	61.6	62
Black	70	19.5	19
Hispanic	59	16.4	16
Asian	*	*	*
Native American	*	*	*
Low income	100	27.9	28
Not low income	259	72.1	72
Individualized education plan	59	16.4	16
No individualized education plan	300	83.6	84
English language learner	40	11.1	11
Not English language learner	319	88.9	89

† Not applicable.

\* Not reported to protect subgroups with fewer than 10 students.

## Recoding the Ends of the Distribution

Another protection implemented by a number of states involves bottom or top coding the results at the tails of the percentage distribution, or for high and low rates. This is typically done by coding all percentages above 95 percent as greater than 95 percent and coding all percentages below 5 percent as less than 5 percent. This is done to avoid reporting the fact that all, or nearly all, of the students in a reporting subgroup share the same achievement level or the same outcome or that very few or none of the students have a particular outcome.

Ideally, this approach is intended to protect categories with 0 to 2 fewer than all students in a reporting category or, conversely, categories with 0 to 2 students. However, with reporting subgroups of 10 to 19 students, all of the percentages of 10 percent or less are based on only 1 student (e.g., 1 of 19 students is 5 percent and 1 of 10 students is 10 percent, while 2 of 19 is 11 percent and 2 of 10 is 20 percent). As a result, with reporting subgroups of 10 to 19 students, even reporting a category as 10 percent or less is no different than reporting that there is at most only 1 student in the category.

The extent of recoding required to protect small categories is related to the size of the subgroup, with a larger recoded range required for smaller subgroups. At a minimum, results should not be published for outcomes based on the experiences of 1 student. The goal is to ensure that each recoded percent could include at least 2 students. Additional protection is provided by including counts of students in the range of recoded percentages where the recoded percent could include at least 3 students (i.e., the threshold rule of 3). For example, in reporting outcome measures for subgroups of 10 to 20, recoding the ends of the distribution to 20 percent or less and 80 percent or more would result in recoding all percentages for categories based on 0 to 2 students (i.e., 20 percent of 10 is 2).<sup>6</sup> In addition, categories of 3 students would be included in the recoded category when there are 15 or more students in the subgroup (i.e., 3 out of 15 is 20 percent).

In reporting outcome measures for groups of 21 to 40, recoding the ends of the distribution to 10 percent or less and 90 percent or more would result in recoding all percentages based

on categories of 0 to 2 students. In this recode, categories of 3 students would be included in the recoded category when there are 30 or more students in the subgroup (i.e., 3 out of 30 is 10 percent).

When there are 41 to 100 students, recoding the ends of the distribution to 5 percent or less and 95 percent or more ensures results based on 0 to 2 students when there are 41 students and 0 to 4 students when there are 100 students (above 59 students, this recode would include categories of 3 students). Similarly, for groups of 101 to 300 students, recoding the ends of the distribution to 2 percent or less and 98 percent or more ensures reporting results based on 0 to 2 students when there are 101 students and 0 to 6 students when there are 300 students (above 149 students this recode includes categories of 3 students). Finally, for groups of more than 300 students, recoding the ends of the distribution to 1 percent or less and 99 percent or more ensures results based on 0 to 3 students at a minimum

Recoding the percentages at one end of a percentage distribution is not necessarily enough to protect the original contents of the recoded category, since the sum of the reported categories subtracted from 100 percent yields the percent that was recoded.

To protect the recoded categories, additional recoding is needed. For groups of 10 to 20 students, the results should be collapsed into two categories and percentages between 21 and 79 should be reported in 10 percentage point ranges. For groups of 21 to 40 students, the percentages in categories of an outcome measure should be recoded in 10 percentage point ranges. For groups of 41 to 200 students, the percentages in categories of an outcome measure should be recoded in 5 percentage point ranges. For groups of 201 or more students, reporting the percentages in categories of an outcome measure as whole numbers provides sufficient recoding (i.e. there are at least 2 counts that could yield each reported percent).

To further protect small categories, if one subgroup includes 200 or fewer students, any related subgroups (i.e., those that combine to sum to the total) with more than 200 students should be recoded using the ranges for 200 students.

---

<sup>6</sup> Reporting results based on fewer than 10 students while ensuring that there could be at least 2 students in a reported category requires more extensive top and bottom coding and would limit the number of reportable outcomes to a small enough set of possible outcomes that they would not be well protected. For example, with results based on 6 students, 2 students account for 33 percent, and recodes of 33 and 67 percent leave only 1 response option that could be reported. Similarly, with 7 students, the recodes would be 29 and 71 percent, leaving 2 response options for reporting; with 8 students, the recodes would be 25 and 75 percent, leaving 3 response options for reporting; and with 9 students, the recodes would be 22 and 78 percent, leaving only 5 response options for reporting.

### ***Example 11: Best Practices: Recoding the Distribution***

Table 14 in this example shows the number of students and the actual and recoded percentage distributions for the school-level third-grade reading assessment results for 32 students for this reporting option. The shaded cells are not publicly reported. Table 14 displays the data with reporting subgroups less than 10 suppressed and the categories of other subgroups recoded to protect small categories. For the overall results of the 32 students, each category is recoded into a 10 percentage point range to protect small categories in the subgroups in the table. Given that there are only 10 students in the Hispanic subgroup, the 0 in the Advanced category is combined with the 10 percent in the proficient category and recoded to less than or equal to ( $\leq$ ) 20 percent at or above proficient, and the 50 percent at the Basic level is combined with the 40 percent at the Below Basic level and recoded to greater than or equal to 80 percent. The data for the 22 White students are recoded, with the 0 percent in the Below Basic category recoded to less than or equal to 10 percent and the other three categories recoded into 10 percentage point ranges. Since there are fewer than 10

students with individualized education plans, the data for this subgroup and the data for students who do not have individualized education plan are suppressed. The outcome measures for the 12 English language learners and the subgroup of 20 students who are not English language learners are reported for those students scoring at the proficient or above level and those performing at or Below the Basic level.

Table 15 follows the same format and shows the results for the district-level third-grade reading assessment results. With 320 students in the group, the results for the 3 students in the advanced category that account for 1 percent of the total are recoded to less than or equal to ( $\leq$ ) 1 percent, and the other three categories are reported as percentages that are rounded to whole numbers. With 198 White students and 122 Hispanic students, the results for the 3 Advanced students in the White subgroup and for 0 Advanced students in the Hispanic subgroup are both recoded to less than or equal to ( $\leq$ ) 2 percent, and the other three categories in each subgroup are recoded into 5 percentage point ranges. With 40

students with individualized education plans, the Advanced category for these students is recoded to less than or equal to ( $\leq$ ) 10 percent, and the remaining categories are recoded into 10 percentage point ranges. The data for the 280 students in the related subgroup who do not have individualized education plans are recoded following the procedures that apply to 200 students, with the 1 percent at the Advanced level recoded to less than or equal to ( $\leq$ ) 2 percent and the other three categories recoded into 5 percentage point ranges. Finally, because there are only 12 students who are English language learners, the Advanced category for these students is combined with the Proficient category and reported as 21 to 29 percent, and the Below Basic and Basic categories are combined and reported as 70 to 79 percent. The data for the 308 students in the related subgroup who are not English language learners are recoded, with the percent at the Advanced level reported as less than or equal to ( $\leq$ ) 2 percent and the other three categories recoded into 5 percentage point ranges.

**Table 14. School-level grade 3 reading assessment results for a state with a minimum reporting size of 10**

		Percent assessed	Tested	Below Basic	Basic	Proficient	Advanced	
<b>Total</b>	N	†	32	4	10	11	7	
	%	100	100	13	31	34	22	Actual
	%	100	100	11–19	30–39	30–39	20–29	Reported
White	N	†	22	0	5	10	7	
	%	100	100	0	23	45	32	Actual
	%	100	100	≤10	21–29	40–49	30–39	Reported
Hispanic	N	†	10	4	5	1	0	
	%	100	100	40	50	10	0	Actual
	%	100	100	†	≥80	≤20	†	Reported
Individualized education plan	N	†	7	4	3	0	0	
	%	100	*	*	*	*	*	<10
	%	100	*	*	*	*	*	<10
No individualized education plan	N	†	25	0	7	11	7	
	%	100	100	0	28	44	28	Actual
	%	100	*	*	*	*	*	Suppressed
English language learner	N	†	12	4	5	2	1	
	%	100	100	33	42	17	8	Actual
	%	100	100	†	70–79	21–29	†	Reported
Not English language learner	N	†	20	0	5	9	6	
	%	100	100	0	25	45	30	Actual
	%	100	100	†	21–29	70–79	†	Reported

† Not applicable.

\* Not reported to protect subgroups with fewer than 10 students.

NOTE: Details may not sum to totals because of rounding and recoding.

**Table 15. District level, Grade 3 reading assessment results for a state with a minimum reporting size of 10**

		Percent assessed	Tested	Below Basic	Basic	Proficient	Advanced	
<b>Total</b>	N		320	40	167	110	3	
	%	100	†	13	52	34	1	Actual
	%	100	†	13	52	34	≤1	Reported
White	N		198	0	105	90	3	
	%	100	†	0	53	45	2	Actual
	%	100	†	≤2	50–54	45–49	≤2	Reported
Hispanic	N		122	40	62	20	0	
	%	100	†	33	51	16	0	Actual
	%	100	†	25–29	50–54	15–19	≤2	Reported
Individualized education plan	N		40	25	15	0	0	
	%	100	†	63	38	0	0	Actual
	%	100	†	60–69	30–39	≤10	≤10	Reported
No individualized education plan	N		280	15	152	110	3	
	%	100	†	5	54	39	1	Actual
	%	100	†	5–9	50–54	35–39	≤2	Reported
English language learner	N		12	4	5	2	1	
	%	100	†	33	42	17	8	Actual
	%	100	†	†	70–79	21–29	†	Reported
Not English language learner	N		308	36	162	108	2	
	%	100	†	12	53	35	1	Actual
	%	100	†	10–14	50–54	35–39	≤2	Reported

† Not applicable.

NOTE: Details may not sum to totals because of rounding and recoding.

## Recommendations

This review and analysis of current reporting practices illustrates that some practices work better than others in protecting suppressed results and, thus, in protecting against disclosures of personally identifiable information about individual students. It is important to note that each of the practices requires some loss of information. The challenge rests in identifying practices that protect information about individual students while minimizing the negative impact on the utility of the publicly reported data. Drawing upon the review and analysis presented in this brief leads to recommended reporting rules to be used in producing reports of percentages and rates to describe student outcomes to the public. These rules are intended for use in the public release of new data.

Rules 1 through 4 and 6 and 7 are general reporting rules. Rule 5 is guided by the number of students in the reporting group or subgroups; the underlying principle is that the amount of detail that can be reported while protecting each

student's privacy is related to the number of students in a reporting group or subgroup—that is, more detail can be reported for larger groups. Rule 5a applies to instances in which there are more than 300 students in each of a set of related reporting subgroups (e.g., in each race/ethnicity group, for students with and without an individualized education plan, for students receiving or not receiving instruction as an English language learner). Rule 5b applies to instances in which the smallest reporting subgroup within a set of related reporting subgroups has 201 to 300 students. Rule 5c applies to instances in which the smallest reporting subgroup within a set of related reporting subgroups has 101 to 200 students. Rule 5d applies when the smallest reporting subgroup in a set of related subgroups has 41 to 100 students. Rule 5e applies when the smallest reporting subgroup in a set of related subgroups has 21 to 40 students. Rule 5f applies when the smallest reporting subgroup in a set of related subgroups has 10 to 20 students.

### Reporting Rules

1. Minimize the amount of enrollment details reported in the profile of the school, district, or state in reports of outcome measure results. If possible, use enrollment data for a different date than that of the reported outcome measures and label the different dates (e.g., report enrollment data for a date different from the assessment date, such as fall enrollment for a spring assessment). In so doing, tell the readers that the data on student enrollment by grade and by selected student characteristics are included to provide context for the results presented but should not be assumed to exactly match the student composition at the time the outcome was measured.
  - a. Report the percentage distribution of students by grade at the school, district, or state level in a standalone table without any of the outcome measures or reporting subgroup details.
  - b. Report the percentage distribution of students by reporting subgroup at the school, district, or state level in a standalone table without any of the outcome measures or enrollment by grade details.
  - c. Do not report the details of the enrollment data within each reporting subgroup by individual grades.
4. Use a minimum of 10 students for the reporting subgroup size limitation.
  - a. Suppress results for all reporting groups with 0 to 9 students.
  - b. Suppress results for reporting subgroups with 0 to 9 students and suppress each of the related reporting subgroups regardless of the number of students in the subgroup (i.e., suppress the other subgroup(s) of the set of subgroups that sum to the overall group). In instances with 3 or more subgroups, the subgroups with 0 to 9 students can be combined with each other or with the smallest reportable subgroup to form an aggregated subgroup of 10 or more students to allow for the reporting of data for larger subgroups.
3. Use only whole numbers when reporting the percentage of students for each category of an outcome measure (e.g., the percentage assessed).



4. Do not report the underlying counts for the subgroup or group totals (i.e., the denominators of the percentages); also do not report the underlying counts of students in individual outcome categories (i.e., the numerators).
5. **To implement the next step in the data protection procedure in the remaining reporting groups and subgroups, the approach used is determined by the number of students in the smallest reporting subgroup among a set of related groups or subgroups (i.e., groups that in combination sum to the total). To protect student privacy:**
  - a. **For reporting variables/outcome measures with more than 300 students and no related subgroup with fewer than 200 students, use the following approach:**
    - i. Recode categories with values of 99 to 100 percent to greater than or equal to 99 percent ( $\geq 99$  percent).
    - ii. Recode categories with values of 0 to 1 percent to less than or equal to 1 percent ( $\leq 1$  percent).
    - iii. Otherwise, report the percentage of students in each category using whole numbers.
  - b. **For reporting variables/outcome measures with 201 to 300 students and no related subgroup with fewer than 200 students, use the following approach:**
    - i. Recode categories with values of 98 to 100 percent to greater than or equal to 98 percent ( $\geq 98$  percent).
    - ii. Recode categories with values of 0 to 2 percent to less than or equal to 2 percent ( $\leq 2$  percent).
    - iii. Otherwise, report the percentage of students in each category using whole numbers.
  - c. **For reporting variables/outcome measures in which the number of students ranges from 101 to 200, use the following option in this group and all related subgroups with more than 200 students:**
    - i. Recode categories with values of 98 to 100 percent to greater than or equal to 98 percent ( $\geq 98$  percent).
    - ii. Recode categories with values of 0 to 2 percent to less than or equal to 20 percent ( $\leq 2$  percent).
    - iii. Recode the percentage in each remaining category in all reporting groups or subgroups to intervals as follows (3–4, 5–9, 10–14, 15–19, . . . , 85–89, 90–94, 95–97).
  - d. **For reporting variables/outcome measures in which the number of students in the smallest reporting group or subgroup ranges from 41 to 100, use the following option in that group or subgroup and use option 5c for each related reporting group or subgroup with more than 100 students:**
    - i. Recode categories with values of 95 to 100 percent to greater than or equal to 95 percent ( $\geq 95$  percent).
    - ii. Recode categories with values of 0 to 5 percent to less than or equal to 5 percent ( $\leq 5$  percent).
    - iii. Recode the percentage in each remaining category in all reporting groups or subgroups to intervals as follows (6–9, 10–14, 15–19, 20–24, . . . , 85–89, 90–94).

- e. **For reporting variables/outcome measures in which the number of students in the smallest reporting group or subgroup ranges from 21 to 40**, use the following option for that group or subgroup, use option 5d for each related reporting group or subgroup with 41 to 100 students, and use option 5c for those with more than 100 students:
  - i. Recode categories with values of 90 to 100 percent to greater than or equal to 90 percent ( $\geq 90$  percent).
  - ii. Recode categories with values of 0 to 10 percent to less than or equal to 10 percent ( $\leq 10$  percent).
  - iii. Recode the percentage in each remaining category in all reporting groups or subgroups to intervals as follows (11–19, 20–29, . . . , 80–89).
- f. **For reporting variables with 10 to 20 students in the smallest subgroup**, use the following option for that group or subgroup, use option 5e for each related group or subgroup with 21 to 40 students, use option 5d for those with 41 to 100 students, and use option 5c for those with more than 100 students:
  - i. Collapse all outcome measures to only two categories, using the same collapsing rules across all subgroups for each outcome measure (e.g., assessment results collapsed to below the proficient level and at or above the proficient level by sex, racial and ethnic groups, disability status, etc.).
- ii. Recode categories with values of 0 to 20 percent to less than or equal to 20 percent ( $\leq 20$  percent), and recode the other category to greater than 80 percent ( $> 80$  percent).
- iii. If both collapsed categories have percents of 21 to 79 percent, recode the percentage in each collapsed category to intervals as follows (21–29, 30–39, . . . , 70–79).
6. For each outcome measure reported at the district level, if results for a group or subgroup have been collapsed, recoded, or suppressed in only one school in the district, apply the same collapsing, recoding, or suppression rule for that group or subgroup in a second school or at the district level (i.e., for any specific measure and group or subgroup, there must be either no school-level data suppressed for a specific subgroup or the data for that subgroup must be suppressed for at least 2 schools or for one school and the district).
7. For each outcome measure reported at the state level, if results for a group or subgroup have been collapsed, recoded, or suppressed in only one district in the state, apply the same collapsing, recoding, or suppression rule for that group or subgroup in a second district (i.e., for any specific measure and group or subgroup, there must be either no district-level data suppressed for a specific subgroup or the data for that subgroup must be recoded or suppressed for at least 2 districts).

## Summary

This Brief discusses the potential for the disclosure of personally identifiable information in summary school-, district-, and state-level reports from education records using current reporting practices. Building on current best practices, the Brief outlines reporting recommendations. Primarily, the goal of these reporting recommendations is to maximize the reporting of student outcomes while protecting students' personally identifiable information.

While it would be easier to have only one set of reporting recommendations, the reporting rules are intended to maximize the amount of detail that can be safely reported without allowing the disclosure of student outcome measure categories based on small numbers of students. A secondary goal of these recommendations is to maximize uniformity in reporting practices across states in order to facilitate cross-state comparisons.

The recommendation to provide data on enrollment by grade and enrollment by student characteristics that are not identical to those for the day the outcome is measured is intended to prevent the statistical manipulation of the data to recover protected student information. However, this may not always be possible, and in some instances, these data may not change over the course of a school year. Thus, the reporting rules

that are linked to the number of students included in a subgroup are intended to add additional protections by ensuring that, if the subgroup size is known, each reported category could include at least two students. Further, if the subgroup size is not known, each reported category could include at least three students.

There are multiple approaches to statistical data protection. The recommendations here were selected with the goal of maximizing the amount of information that can be released while protecting personally identifiable student information through a relatively straightforward set of rules that can be easily implemented. For those readers wanting to read further on the topic of statistical data protection, please see Duncan et. al. (1993) *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*; Willenborg and de Waal (2001) *Statistical Disclosure Control in Practice*; Federal Committee on Statistical Methodology Working Paper 22, *Report on Statistical Disclosure Limitation Methodology*; and the American Statistical Association, Committee on Privacy and Confidentiality website, *Key Terms/Definitions in Privacy and Confidentiality*.

NCES welcomes input on these recommendations.

## References

- American Statistical Association, Committee on Privacy and Confidentiality. *Key Terms/Definitions in Privacy and Confidentiality*. Alexandria, VA: Retrieved from <http://www.amstat.org/committees/pc/keyterms.html> on 6/17/2010.
- Code of Federal Regulations, Title 34—Education, Part 200. *Improving the Academic Achievement of the Disadvantaged, Section 200.7, Disaggregation of Data*, (34CFR200.7). Washington, DC: GPO Access CFR. Retrieved from [http://edocket.access.gpo.gov/cfr\\_2010/julqtr/34cfr200.7.htm](http://edocket.access.gpo.gov/cfr_2010/julqtr/34cfr200.7.htm) on 10/10/2010.
- Code of Federal Regulations, Title 34—Education, Part 99. *Family Educational and Privacy Rights*, (34CFR99). Washington, DC: GPO Access e-CFR. Retrieved from [http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&sid=44d350c26fb9c4a156bf805f297c9e&tpl=/ecfrbrowse/Title34/34cfr99\\_main\\_02.tpl](http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&sid=44d350c26fb9c4a156bf805f297c9e&tpl=/ecfrbrowse/Title34/34cfr99_main_02.tpl).
- Duncan, George T., Jabine, Thomas B. and de Wolf, Virginia A., Editors. (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Panel on Confidentiality and Data Access, National Research Council. Washington, DC: National Academy Press.
- Federal Register, Office of Management and Budget, *Implementation Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)*, Washington DC: Vol. 72, No. 115 / Friday, June 15, 2007. Retrieved from [http://www.whitehouse.gov/sites/default/files/omb/assets/omb/fedreg/2007/061507\\_cipsea\\_guidance.pdf](http://www.whitehouse.gov/sites/default/files/omb/assets/omb/fedreg/2007/061507_cipsea_guidance.pdf) on 9/9/2010.
- Office of Management and Budget, Federal Committee on Statistical Methodology, (2005). Statistical Policy Working Paper 22, *Report on Statistical Disclosure Limitation Methodology*.

Retrieved from <http://www.fcsn.gov/workingpapers/spwp22.html> on 9/9/2010.

Office of Management and Budget, OMB  
Memorandum M-03-22, *OMB Guidance for Implementing the Privacy Provisions of the E-Government Act of 2002*. Retrieved from <http://www.whitehouse.gov/omb/memoranda/m03-22/> on 9/9/2010.

Office of Management and Budget, OMB  
Memorandum M-07-16, *Safeguarding Against and Responding to the Breach of Personally Identifiable Information*. Retrieved from <http://www.whitehouse.gov/sites/default/files/omb/assets/omb/memoranda/fy2007/m07-16.pdf> on 9/9/2010.

U.S. Code, Title 20—Education, Chapter 31—General Provisions Concerning Education, Subchapter III—General Requirements and Conditions Concerning Operation and Administration of Education Programs: General Authority of Secretary, Part 4—Records, Privacy, Limitation on Withholding Federal funds, Section 1232g. *Family Educational and Privacy Rights*, (20USC1232g). Washington, DC: GPO Access. Retrieved from <http://frwebgate4.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=799486197532+0+1+0&WAIAction=retrieve>.

U.S. Code, Title 20—Education, Chapter 70—Strengthening and Improvement of Elementary and Secondary Schools, Subchapter I—Improving the Academic Achievement of the Disadvantaged, Part A—Improving Basic Programs Operated by Local Educational

Agencies, Subpart 1—Basic Program Requirements, Section 6311. *State Plans*, (20USC6311). Washington, DC: GPO Access. Retrieved from <http://frwebgate2.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=bULwJH/21/1/0&WAIAction=retrieve>.

U.S. Code, Title 20—Education, Chapter 76—Education Research, Statistics, Evaluation, Information, and Dissemination, Subchapter II—Educational Technical Assistance, Section 9607. *Grant Program for Statewide, Longitudinal Data Systems*, (20USC9607). Washington, DC: GPO Access. Retrieved from <http://frwebgate3.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=FKr6BA/0/1/0&WAIAction=retrieve> on 9/9/2010.

Public Law 110-69, America Competes Act, Title VI—Education, Section 6401. Washington, DC: GPO Access. Retrieved from [http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110\\_cong\\_public\\_laws&docid=f:publ069.110](http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110_cong_public_laws&docid=f:publ069.110) on 9/10/2010.

Public Law, 111-05, American Recovery and Reinvestment Act, Title VIII—Education, Institute of Education Sciences. Washington, DC: GPO Access. Retrieved from [http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=111\\_cong\\_public\\_laws&docid=f:publ005.111](http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=111_cong_public_laws&docid=f:publ005.111) on 9/10/2010.

Willenborg, L., and De Waal, T. (2001). *Elements of Statistical Disclosure Control*, Vol. 155, Lecture Notes in Statistics, New York, NY: Springer.



ALLIANCE FOR  
EXCELLENT EDUCATION

# Ensuring Equity in ESSA:

## The Role of N-Size in Subgroup Accountability

June 2016



# Table of Contents

What Is N-Size and Why Does It Matter? . . . . . 3

Consistency and Comparability of Data . . . . . 4

Protecting Student Privacy and Ensuring Statistical Reliability . . . . . 7

Strengthening Student Subgroup Accountability . . . . . 8

Policy Recommendations . . . . . 8

Conclusion . . . . . 9

Closing Achievement Gaps with Evidence-Based Reform and Interventions . . . . 10

Endnotes . . . . . 11

# Abstract

States are responsible for setting the minimum number of students needed to form a student subgroup for federal reporting and accountability purposes. This required student subgroup size is commonly referred to as the state-set “n-size.” States should set this number as low as possible to maximize the number of student subgroups created. This will ensure that states identify student subgroups with low academic performance and/or low high school graduation rates and provide targeted interventions to support the schools those students attend. Specifically, states should not require a subgroup to include more than ten students to include that subgroup for reporting and accountability purposes.

# Acknowledgments

*This paper was written by **Jessica Cardichon, EdD**, senior director of policy and advocacy for comprehensive high school reform at the Alliance for Excellent Education (the Alliance). **Sean Bradley**, policy and advocacy intern at the Alliance, also contributed to this paper. **Aharon Charnov**, website and video production manager at the Alliance, designed the cover art for this paper.*

*The Alliance acknowledges the **Bill & Melinda Gates Foundation** and **Carnegie Corporation of New York** for their generous financial support for the development of this paper. The findings and opinions expressed are those of the Alliance and do not necessarily represent the views of the Bill & Melinda Gates Foundation or Carnegie Corporation of New York.*

*The **Alliance for Excellent Education** is a Washington, DC–based national policy and advocacy organization dedicated to ensuring that all students, particularly those traditionally underserved, graduate from high school ready for success in college, work, and citizenship. [www.all4ed.org](http://www.all4ed.org)*

© Alliance for Excellent Education, June 2016.



## What Is “N-Size” and Why Does It Matter?

At its core, the Elementary and Secondary Education Act (ESEA) is a civil rights law with the primary purpose of ensuring that historically underserved students have equitable access to the educational opportunities they need to reach their full potential. Knowing the achievement level of individual students is fundamental to knowing whether the purpose of this law is being fulfilled.



During its time, the No Child Left Behind Act (NCLB), the previous bill to reauthorize ESEA, required states to report on the performance of historically underserved students—including students of color, students from low-income families, and students with disabilities—and held them accountable for gaps in performance. While NCLB’s approach to addressing those performance gaps was misguided, its requirement to reveal how these students were performing was a critical first step to ensuring equity.

Prior to NCLB, the overall performance of a school often masked the performance of student subgroups, hiding gaps in academic achievement and high school graduation rates for historically underserved students.<sup>1</sup> The recently passed Every Student Succeeds Act (ESSA) of 2015 requires states, districts, and schools

to identify low-performing subgroups of students, report on their progress, and provide targeted intervention and support when they consistently demonstrate low performance.

The key term in this requirement is “subgroups” of students, which refers to student groups based on racial/ethnic status, socioeconomic status, English-language ability, and disability status. Under ESSA, as under NCLB, states set the minimum number of students required to create a subgroup of students at the school, district, and state levels. This state-set number, commonly referred to as the “n-size,” must not reveal personally identifiable information about the student and must yield statistically reliable information.<sup>2</sup> However, a significant number of states set their n-size higher than necessary to meet the requirements originally set under NCLB and maintained under ESSA.

Additionally, setting the n-size too high interferes with a state's ability to meet the student subgroup accountability requirements<sup>3</sup> under ESSA. ESSA requires states to identify schools with consistently underperforming subgroups of students and implement evidence-based, targeted intervention in these schools.

However, if a school does not have enough students from a particular subgroup to reach the state-set n-size, then the school does not have to report the academic performance or high school graduation rates of students in that subgroup and ESSA does not require interventions and support for those students. For example, if a state sets the n-size at 30 students and a high school has only twenty-nine African American students in the twelfth-grade class, that subgroup of African American students essentially does not exist for reporting and accountability purposes. The individual students would count in the high school's overall graduation rate, but the school would not report any gaps between the graduation rate of African American students and their white peers in that particular high school, nor would the school receive any intervention and support to address those gaps.

If states set the n-size higher than necessary to be statistically sound and protect student privacy, they are *less likely* to reveal the low performance of student subgroups. Consequently, they are *more likely* to overlook a number of student subgroups for both reporting and accountability purposes and underidentify schools needing and receiving targeted intervention and support.




## Consistency and Comparability of Data


Consistency across states in terms of comparable data is also an important goal to ensure accurate cross-state comparisons of gaps in student subgroup performance. Currently, significant variation exists across states regarding the minimum number of students needed for a student subgroup to exist for federal reporting and accountability and improvement purposes. As table 1 shows, for federal accountability and improvement purposes


- thirteen states set an n-size of 10 or fewer students;<sup>4</sup>
- nine states and California's [CORE Districts](#)<sup>5</sup> set the n-size between 11 and 20 students;<sup>6</sup> and
- twenty-eight states and the District of Columbia set the n-size at 21 or more students<sup>7</sup> (eight of those states set it at 31 or more students<sup>8</sup>).





**TABLE 1: State N-Size**
 States with N-Size of 10 or Less

 States with N-Size Between 11 and 20

 States with N-Size of 21 or More

STATE	N-Size for Federal Accountability and Improvement Purposes	N-Size for Reporting Academic Performance and High School Graduation Rates
Alaska <sup>9</sup>	5	5
Maryland <sup>10</sup>	5	5
Wyoming <sup>11</sup>	6	6
Florida <sup>12</sup>	10	10
Iowa <sup>13</sup>	10	10
Maine <sup>14</sup>	10	10
Mississippi <sup>15</sup>	10	10
Nebraska <sup>16</sup>	10	10
North Dakota <sup>17</sup>	10	10
Oklahoma <sup>18</sup>	10	10
South Dakota <sup>19</sup>	10	10
Utah <sup>20</sup>	10	10
West Virginia <sup>21</sup>	10	10
New Hampshire <sup>22</sup>	11	11
Georgia <sup>23</sup>	15	10
Alabama <sup>24</sup>	20	10
Colorado <sup>25</sup>	16/20 <sup>a</sup>	No minimum set
Connecticut <sup>26</sup>	20	20
CORE Districts (California) <sup>27</sup>	20	20
Massachusetts <sup>28</sup>	20	6/10 <sup>b</sup>
Minnesota <sup>29</sup>	20	10
Rhode Island <sup>30</sup>	20	20
Wisconsin <sup>31</sup>	20	20
Arkansas <sup>32</sup>	25	25
District of Columbia <sup>33</sup>	25	10
Idaho <sup>34</sup>	25	25
Kentucky <sup>35</sup>	25/10 <sup>c</sup>	10

*(continued)*

**Notes:** N-size refers to the minimum number of students needed within a specific subgroup to create that subgroup for federal reporting and accountability purposes.

<sup>a</sup> Colorado uses an n-size of 16 students for the academic achievement and high school graduation rates of student subgroups and an n-size of 20 students for growth in academic achievement for student subgroups.

<sup>b</sup> Massachusetts uses an n-size of 10 students for reporting the academic performance of student subgroups and 6 students for reporting high school graduation rates of student subgroups on school report cards.

<sup>c</sup> Kentucky uses an n-size of 25 students to identify the bottom 5 percent of student subgroups and an n-size of 10 students for the "nonduplicated student gap group." See Kentucky Department of Education, ESEA Flexibility Request, <http://www2.ed.gov/policy/eseaflex/approved-requests/ky3req32015.doc>.

**TABLE 1: State N-Size** *(continued)*

States with N-Size of 10 or Less

States with N-Size Between 11 and 20

States with N-Size of 21 or More

STATE	N-Size for Federal Accountability and Improvement Purposes	N-Size for Reporting Academic Performance and High School Graduation Rates
Nevada <sup>36</sup>	25	10
Texas <sup>37</sup>	25	25
Delaware <sup>38</sup>	30	30
Indiana <sup>39</sup>	30	10
Kansas <sup>40</sup>	30	30
Michigan <sup>41</sup>	30	30
Missouri <sup>42</sup>	30	30
Montana <sup>43</sup>	30/10 <sup>d</sup>	6
New Jersey <sup>44</sup>	30	10
New York <sup>45</sup>	30	5
North Carolina <sup>46</sup>	30	10
Ohio <sup>47</sup>	30	30
Pennsylvania <sup>48</sup>	30	30
South Carolina <sup>49</sup>	30	30
Tennessee <sup>50</sup>	30	10
Virginia <sup>51</sup>	30	30
Washington <sup>52</sup>	30	10
Arizona <sup>53</sup>	40	10
Hawaii <sup>54</sup>	40	40
Louisiana <sup>55</sup>	40/10 <sup>e</sup>	10
New Mexico <sup>56</sup>	40	10
Oregon <sup>57</sup>	40/30/20 <sup>f</sup>	40/30/20
Vermont <sup>58</sup>	40	11
Illinois <sup>59</sup>	45	10
California <sup>60</sup>	50	50

**Notes:** N-size refers to the minimum number of students needed within a specific subgroup to create that subgroup for federal reporting and accountability purposes.

<sup>d</sup> Montana uses an n-size of 30 students for federal accountability purposes. For small schools that test fewer than thirty students overall, which account for approximately 40 percent of the state's schools, Montana uses an n-size of 10 students for federal accountability purposes.

<sup>e</sup> Louisiana uses an n-size of 40 students for high school graduation rates and an n-size of 10 students for performance on assessments for federal accountability purposes.

<sup>f</sup> Oregon uses an n-size of 30 students for the overall growth in student academic achievement and the growth in academic achievement for student subgroups and an n-size of 40 students for the overall high school graduation rate and student subgroup high school graduation rate. However, Oregon uses two years of data when reporting student performance and high school graduation rates and uses four years of data for small schools. So while 40 students is the minimum n-size for reporting high school graduation rates, this is forty students over two consecutive cohorts combined. This means that each student subgroup cohort must average twenty students per year (and only ten students per year in small schools) to be included for federal accountability purposes.

This extreme variation makes cross-state comparisons of student subgroup performance difficult. For example, Maryland currently has an n-size of 5 students, while Louisiana has an n-size of 40 students. The National Center for Education Statistics (NCES) notes that setting a maximum n-size that allows for less varying extremes creates greater “uniformity in reporting practices across states in order to facilitate cross-state comparisons.”<sup>61</sup> Further, when states set an unnecessarily high n-size, it increases the likelihood that they will underreport the number of schools with gaps in the performance of student subgroups, limiting their ability to provide additional support to a significant number of historically underserved students.

Additionally, the U.S. Department of Education’s (ED’s) Office of Special Education and Rehabilitative Services (OSERS) recommends that states set a consistent n-size of 10 for the purpose of determining whether “significant disproportionality” exists among racial/ethnic groups in the rates at which students with disabilities within each racial/ethnic group are disciplined.<sup>62</sup> According to the proposed rules from OSERS, wide variations exist across states in the n-size they use to create the racial/ethnic groups to determine whether students with disabilities within these groups are disciplined at varying rates based on race. For this purpose, nine states set the n-size at 10 students, while four states set the n-size at 30 students, for example. If a school does not have enough students from a particular racial/ethnic subgroup to reach the n-size, then the school does not have to examine whether students with disabilities within that racial/ethnic group are disciplined at disproportionate rates.

ED notes that when states set a higher n-size, they eliminate more student subgroups, and school districts, from the analysis, thereby limiting the number of students states can identify for additional support. When states set an unnecessarily high n-size for the purpose of determining “significant disproportionality” they undermine accountability in the same way that high n-sizes undermine ESSA’s reporting and accountability provisions. ED proposes setting the maximum n-size at 10 students to address these concerns and “ensure that States examine as many racial and ethnic groups for significant disproportionality in as many [districts] as possible,” according to the proposed rules.<sup>63</sup>

## Protecting Student Privacy and Ensuring Statistical Reliability

Under the Family Educational Rights and Privacy Act,<sup>64</sup> state reporting of disaggregated student data, such as student subgroups, may not be published if the results would yield personally identifiable information<sup>65</sup> about an individual student. In addition, ESSA requires<sup>66</sup> states to set an n-size that protects student privacy and is sufficient to yield statistically reliable information. According to a report by NCES,<sup>67</sup> a state can set an n-size of 10 students, and even as low as 5 students, and fully meet the requirement for statistical reliability and also fully protect student privacy. The NCES report also describes several statistical methods states are using to protect student privacy. For example, some states use “various forms of [data] suppression, top and bottom coding of values at the ends of a [data] distribution, and limiting the amount of detail reported for the underlying [number of students]” to provide statistically reliable information that protects individual student privacy.<sup>68</sup>



## Strengthening Student Subgroup Accountability

A number of states have demonstrated that by lowering their n-size, they are able to identify and support substantially more schools and students:

- Massachusetts was able to include 100 additional schools in its system of school accountability and support by lowering its n-size from 40 to 30 students.<sup>69</sup>
- The California CORE Districts chose to use an n-size of 20 students, which is lower than the state-set n-size of 50 students and, collectively, were able to include 150,000 additional students in their accountability and support systems.<sup>70</sup>
- Mississippi lowered its n-size from 40 to 30 students and the number of schools accountable for students with disabilities increased from 234 to 872. Similarly, the number of schools accountable for English language learners increased from 15 to 447.<sup>71</sup>
- Virginia lowered its n-size from 50 to 30 students. Consequently, the approximate number of schools accountable for African American students increased from 353 to 451 and those accountable for Latino students increased from 122 to 183. The number of schools accountable for students with disabilities increased from 105 to 396, for English language learners from 104 to 139, and for students eligible for free or reduced-price lunch from 672 to 717.<sup>72</sup>
- Sixteen states and the CORE Districts in California lowered their n-sizes within the last two years.<sup>73</sup>

More states should follow these examples and structure their accountability and support systems to expand, rather than limit, the number of student subgroups included within those systems.



## Policy Recommendations

### Federal Recommendations

ED should issue regulations under ESSA that prohibit states from setting an n-size above 10 students for reporting and accountability purposes unless the state demonstrates that setting a higher number would not exclude a significant number of students and schools. Under this regulation, states still would maintain the flexibility to set an n-size below 10 students.

ED has the authority to place these parameters around the state determination of n-size to ensure that states meet reporting and accountability requirements under ESSA. Although under ESSA,<sup>74</sup> the U.S. Secretary of Education is prohibited from setting a *minimum* number of students needed to form a subgroup, there is no language within ESSA prohibiting the Secretary from setting a *maximum* n-size or a cap.

The Secretary has the authority to ensure that states meet subgroup accountability requirements. In addition, more accurate cross-state comparisons can be made when there is less variation in state-set n-sizes. Further, this would allow for consistency with the maximum n-size that OSERS proposes.



## State and Local Recommendations

As states consider changes to their accountability and improvement systems, they should set their n-size at 10 or fewer students to ensure they capture the greatest number of student subgroups for reporting, accountability, and improvement purposes under ESSA. When states include these schools in their accountability and improvement systems, the schools become eligible for school improvement funding and direct student services under the law. In addition, states may choose to target other federal and state resources to these schools, such as professional development funding under Title II of ESSA. States and districts should prioritize schools with the greatest numbers and percentages of low-performing students as measured by student achievement and high school graduation rates.

There are a number of evidence-based interventions and strategies that these schools can implement to help close gaps in achievement and high school graduation rates including personalization, early-warning identification and intervention systems, and expanded access to rigorous and advanced course work, among others. (See the sidebar on the next page, "Closing Achievement Gaps with Evidence-Based Interventions," for additional information and examples.)

## Conclusion

The ability of state and school accountability systems to identify and support student subgroups inherently depends upon the existence of those individual subgroups within a state's accountability system. States must accurately determine and report the performance of *all* student subgroups in order to thoroughly identify gaps in student performance, prioritize and target resources, and ensure that the schools serving these students receive the support they need to help close these gaps.

An n-size set higher than necessary to protect student information and be statistically sound is counterproductive to identifying and closing those gaps. The promise of ESSA to ensure that every student succeeds will never be fulfilled unless states structure their accountability and improvement systems to be as inclusive as possible. By setting an n-size of 10 or fewer students, state accountability systems effectively can identify and support the nation's underserved students and realize the civil rights imperative inherent within the law.



## Closing Achievement Gaps with Evidence-Based Reform and Interventions

### Personalization

MDRC conducted an evaluation<sup>75</sup> of New York City's "small schools of choice," which implemented a number of strategies, including an increased focus on personalization. As a result of these reform efforts, the overall high school graduation rates have increased from 60.9 percent to 70.4 percent—9.5 percentage points overall; 13.5 percentage points for African American males and 10.3 percentage points for Latino females.<sup>76</sup> The increase in four-year high school graduation rates is equivalent to nearly half of the gap in graduation rates between white students and students of color in New York City. In addition, this initiative has led to an overall increase in college enrollment of 8 percentage points and an increase in college enrollment for African American males of 11 percentage points, a 36 percent increase relative to their peers.<sup>77</sup> Principals and teachers at these schools with the strongest evidence of effectiveness strongly believe that academic rigor and personal relationships account for the effectiveness of their schools.

The Chicago Public School System effectively uses data to provide students with personalized intervention and support. In Chicago, the city's high school graduation rate rose from 47 percent in 1999 to 69 percent in 2013. This progress resulted from a focused effort to keep Chicago's ninth-grade students on track toward graduation by using data to individualize instruction. The University of Chicago Urban Education Institute predicts that Chicago's graduation rate will exceed 80 percent within the next few years.<sup>78</sup>

### Early-Warning Identification and Intervention Systems

Early-warning identification and intervention systems are based on a broad body of research supporting their use in secondary schools. For example, Diplomas Now partners with the school community and works with administrators and teachers to improve student attendance, behavior, and course performance. They develop a strategic plan, implement an early-warning system to identify struggling students, and regularly review data to foster continuous improvement. For these students, Diplomas Now provides additional academic support in areas of identified need and forms support groups and connects them with community resources, such as counseling, health care, housing, food, and clothing.<sup>79</sup> MDRC recently conducted a first-year process evaluation<sup>80</sup> of Diplomas Now and reports impressive results. For School Year 2013–14, Diplomas Now reports a 62 percent reduction in student suspension, a 58 percent reduction in students failing English, and a 54 percent reduction in students failing math.

### Advanced Placement and International Baccalaureate Programs

Research demonstrates that Advanced Placement (AP) students are more likely to enroll in a four-year college, perform better in college, return for a second year in college, and graduate from college than their non-AP peers.<sup>81</sup> Students—including women and underrepresented students—who take AP math or science exams are more likely to major in STEM (science, technology, mathematics, and engineering) fields.<sup>82</sup> Further, a recent study on students completing the International Baccalaureate (IB) program demonstrates postsecondary education outcomes for students from low-income families. Specifically, students from Title I schools in the IB Diploma Program (DP) enroll in college at the same rate as IB DP students from public schools generally, a rate of 82 percent.<sup>83</sup> Further, IB DP students from low-income families enroll in postsecondary education at a rate of 79 percent compared to the national average for students from low-income families, which is 46 percent.<sup>84</sup>

### Early College/Dual-Enrollment Programs

Research shows that participation in dual-enrollment courses, which allow students to earn high school and college credit simultaneously, can increase high school graduation rates and increase college enrollment and persistence. In [early college high schools](#), where students can earn both a high school diploma and an associate's degree or up to two years of credit toward a bachelor's degree, 90 percent of students graduate from high school and 30 percent earn an associate's degree or other postsecondary credential while in high school.<sup>85</sup>

### Linked Learning

[Linked Learning](#) is an approach to high school redesign being implemented in California that integrates rigorous academics, career-based learning in the classroom, work-based learning in professional settings, and integrated student supports. [Research](#) from SRI International assessing the effect of Linked Learning on students' high school outcomes finds that students enrolled in high-quality Linked Learning pathways are more likely to graduate from high school than other students.

## Endnotes

- <sup>1</sup> A. Williams et al., *State Education Indicators with a Focus on Title I 2000–01* (DOC# 2004-17) (Washington, DC: U.S. Department of Education, 2004), <http://www2.ed.gov/rschstat/eval/disadv/indicators-2000-01/final-report.pdf> (accessed March 30, 2016).
- <sup>2</sup> See ESSA, section 1111 (c)(3)(A).
- <sup>3</sup> See ESSA, sections 1111(c)(4)(D)(III) and (d)(3)(A)(i)(II).
- <sup>4</sup> The states are Alaska, Florida, Iowa, Maine, Maryland, Mississippi, Nebraska, North Dakota, Oklahoma, South Dakota, Utah, West Virginia, and Wyoming.
- <sup>5</sup> CORE represents nine member school districts in California, including Fresno, Garden Grove, Long Beach, Los Angeles, Oakland, Sacramento, San Francisco, Sanger, and Santa Ana Unified. Combined, these districts serve more than 1 million students.
- <sup>6</sup> The states are Alabama, Colorado, Connecticut, Georgia, Massachusetts, Minnesota, New Hampshire, Rhode Island, and Wisconsin.
- <sup>7</sup> The states are Arizona, Arkansas, California, Delaware, Hawaii, Idaho, Illinois, Indiana, Kansas, Kentucky, Louisiana, Michigan, Missouri, Montana, Nevada, New Jersey, New Mexico, New York, North Carolina, Ohio, Oregon, Pennsylvania, South Carolina, Tennessee, Texas, Vermont, Virginia, and Washington.
- <sup>8</sup> The states are Arizona, California, Hawaii, Illinois, Louisiana, New Mexico, Oregon (for the overall high school graduation rate and student subgroup high school graduation rate), and Vermont.
- <sup>9</sup> Alaska Department of Education & Early Development, ESEA Flexibility Request, p. 69, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/akrenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>10</sup> Maryland State Department of Education, ESEA Flexibility Request, p. 22, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/mdrenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>11</sup> Julie Magee, Wyoming Department of Education, e-mail message to Alliance for Excellent Education, February 19, 2016.
- <sup>12</sup> Florida Department of Education, ESEA Flexibility Request, p. 64–65, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/flrenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>13</sup> Ryan Wise, Iowa Department of Education, e-mail message to Alliance for Excellent Education, April 4, 2016.
- <sup>14</sup> Maine Department of Education, ESEA Flexibility Request, p. 73, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/merenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>15</sup> Mississippi Department of Education, ESEA Flexibility Request, p. 56, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/msrenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>16</sup> Jill Aurand, Nebraska Department of Education, e-mail message to Alliance for Excellent Education, March 21, 2016.
- <sup>17</sup> Frank Snow, North Dakota Department of Public Instruction, e-mail message to Alliance for Excellent Education, March 8, 2016.
- <sup>18</sup> Oklahoma State Department of Education, ESEA Flexibility Request, p. 57, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/okrenewalreq7282015.pdf> (accessed March 3, 2016).
- <sup>19</sup> South Dakota Department of Education, ESEA Flexibility Request, p. 70, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/sdrenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>20</sup> Utah State Office of Education, ESEA Flexibility Request, p. 41, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/utrenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>21</sup> Michele Blatt, West Virginia Department of Education, e-mail message to Alliance for Excellent Education, April 1, 2016.
- <sup>22</sup> New Hampshire Department of Education, ESEA Flexibility Request, p. 63, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/nhrenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>23</sup> Allison Timberlake, Georgia Department of Education, e-mail message to Alliance for Excellent Education, May 27, 2016.
- <sup>24</sup> Melinda Maddox, Alabama State Department of Education, e-mail message to Alliance for Excellent Education, March 29, 2016.
- <sup>25</sup> Colorado Department of Education, ESEA Flexibility Request, p. 276, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/coflexrenewal1192015.pdf> (accessed March 3, 2016).
- <sup>26</sup> Connecticut State Department of Education, ESEA Flexibility Request, p. 67, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/ctrenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>27</sup> California Office to Reform Education, ESEA Flexibility Request, p. 6, <https://www.csba.org/~media/CSBA/Files/GovernanceResources/EffectiveGovernance/201305COREWaiverOverviewMay18.ashxPag> (accessed April 6, 2016).
- <sup>28</sup> Matthew Pakos, Massachusetts Department of Elementary and Secondary Education, e-mail message to Alliance for Excellent Education, March 31, 2016.
- <sup>29</sup> Stephanie Graff, Minnesota Department of Education, e-mail message to Alliance for Excellent Education, March 29, 2016.
- <sup>30</sup> Rhode Island Department of Education, ESEA Flexibility Request, p. 62, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/rirenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>31</sup> Wisconsin Department of Public Instruction, ESEA Flexibility Request, p. 56, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/wirenewalreq15.pdf> (accessed March 3, 2016).
- <sup>32</sup> Arkansas Department of Education, ESEA Flexibility Request, p. 79, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/arrenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>33</sup> Etai Mizrav, District of Columbia Office of the State Superintendent of Education, e-mail message to Alliance for Excellent Education, April 4, 2016.
- <sup>34</sup> Idaho State Department of Education, ESEA Flexibility Request, p. 46, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/idrenewalreq2015.pdf> (accessed March 3, 2016).

- <sup>35</sup> Mary Ann Miller, Kentucky Department of Education, e-mail message to Alliance for Excellent Education, May 26, 2016.
- <sup>36</sup> Russ Keglovits, Nevada Department of Education, e-mail message to Alliance for Excellent Education, May 26, 2016.
- <sup>37</sup> Texas Education Agency, ESEA Flexibility Request, p. 34, <http://www2.ed.gov/policy/eseaflex/approved-requests/txrenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>38</sup> Delaware Department of Education, ESEA Flexibility Request, p. 79, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/dernewalreq2015.pdf> (accessed March 3, 2016).
- <sup>39</sup> Jeff J. Coyne, Indiana Department of Education, e-mail message to Alliance for Excellent Education, May 26, 2016.
- <sup>40</sup> Kansas State Department of Education, ESEA Flexibility Request, p. 105, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/ksrenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>41</sup> Michigan Department of Education, ESEA Flexibility Request, p. 101, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/mirenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>42</sup> Missouri Department of Elementary and Secondary Education, ESEA Flexibility Request, p. 56, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/morenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>43</sup> Scott Furois, Montana Office of Public Instruction, e-mail message to Alliance for Excellent Education, March 30, 2016.
- <sup>44</sup> Jill Hulnick, State of New Jersey Department of Education, e-mail message to Alliance for Excellent Education, March 31, 2016.
- <sup>45</sup> Ira Schwartz, New York State Education Department, e-mail message to Alliance for Excellent Education, March 31, 2016.
- <sup>46</sup> Louis M. Fabrizio, North Carolina Department of Public Instruction, e-mail message to Alliance for Excellent Education, May 26, 2016.
- <sup>47</sup> Ohio Department of Education, ESEA Flexibility Request, p. 78, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/ohrenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>48</sup> Pennsylvania Department of Education, ESEA Flexibility Request, p. 53, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/parenewalreq2015.pdf> (accessed March 3, 2016).
- <sup>49</sup> South Carolina Department of Education, ESEA Flexibility Request, p. 73, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/screnewalreq2015.pdf> (accessed March 3, 2016).
- <sup>50</sup> Candice McQueen, Tennessee Department of Education, e-mail message to Alliance for Excellent Education, April 4, 2016.
- <sup>51</sup> Virginia Department of Education, ESEA Flexibility Request, p. 73, <http://www2.ed.gov/policy/eseaflex/approved-requests/va4req32015.pdf> (accessed March 3, 2016).
- <sup>52</sup> Gil Mendoza, State of Washington Office of Superintendent of Public Instruction, e-mail message to Alliance for Excellent Education, March 30, 2016.
- <sup>53</sup> Cecilia E. Johnson, Arizona Department of Education, e-mail message to Alliance for Excellent Education, March 30, 2016.
- <sup>54</sup> Glenn Nochi, Hawaii State Department of Education, e-mail message to Alliance for Excellent Education, February 16, 2016.
- <sup>55</sup> Jennifer Baird, Louisiana Department of Education, e-mail message to Alliance for Excellent Education, May 26, 2016.
- <sup>56</sup> Cindy Gregory, New Mexico Public Education Department, e-mail message to Alliance for Excellent Education, February 23, 2016.
- <sup>57</sup> Jason Wiens, Oregon Department of Education, e-mail message to Alliance for Excellent Education, March 7, 2016.
- <sup>58</sup> Amy Fowler, Vermont Agency of Education, e-mail message to Alliance for Excellent Education, April 19, 2016.
- <sup>59</sup> Angela Chamness, Illinois State Board of Education e-mail message to Alliance for Excellent Education, April 5, 2016.
- <sup>60</sup> Cheryl Haviland, California Department of Education, e-mail message to Alliance for Excellent Education, March 22, 2016.
- <sup>61</sup> U.S. Department of Education, National Center for Education Statistics, *Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting* (NCES 2011–603), <https://nces.ed.gov/pubs2011/2011603.pdf> (accessed March 21, 2016).
- <sup>62</sup> U.S. Department of Education, "Assistance to States for the Education of Children With Disabilities; Preschool Grants for Children With Disabilities; Proposed Rules," 34 C.F.R. 300, Vol. 81, No. 41, March 2, 2016, <https://www.gpo.gov/fdsys/pkg/FR-2016-03-02/pdf/2016-03938.pdf> (accessed March 22, 2016).
- <sup>63</sup> Ibid.
- <sup>64</sup> 20 U.S.C. § 1232g; 34 C.F.R. 99.
- <sup>65</sup> Personally identifiable information includes but is not limited to the following: (1) student's name; (2) name of the student's parent or other family members; (3) address of the student or student's family; (4) a personal identifier, such as the student's social security number, student number, or biometric record; (5) other indirect identifiers, such as the student's date of birth, place of birth, and mother's maiden name; (6) 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; (7) information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates (34 C.F.R. § 99.3).
- <sup>66</sup> See ESSA, section 1111(c)(3)(A)(iii).
- <sup>67</sup> National Center for Education Statistics, *Statistical Methods*.
- <sup>68</sup> Ibid. The report identifies "best practices" to avoid the unintended disclosure of personally identifiable information, including publishing the percentage distribution across categories of outcome measures with no underlying counts or totals; publishing a collapsed percentage distribution across categories of outcome measures with no underlying counts or totals; publishing counts but using complementary suppression at the subgroup level when a small subgroup is suppressed; limiting the amount of detail published for school background information; recoding the ends of percentage distributions; and recoding high and low rates. These recommendations "were selected with the goal of maximizing the amount of information that can be released while protecting personally identifiable student information through a relatively straightforward set of rules that can be easily implemented."



- <sup>69</sup> Massachusetts Department of Elementary and Secondary Education, ESEA Flexibility Request, p. 55, <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/marenewalreq2015.pdf> (accessed April 20, 2016).
- <sup>70</sup> California CORE Districts, "The School Quality Improvement Index & the CORE Data Collaborative," PowerPoint presentation, January 11, 2016, <http://coredistricts.org/wp-content/uploads/2016/01/CORE-Data-Collaborative-v3-1-21-16.pdf> (accessed April 20, 2016).
- <sup>71</sup> Mississippi Department of Education, "Highlights of Mississippi's ESEA Flexibility Request," <https://www2.ed.gov/policy/eseaflex/highlights/ms.doc> (accessed April 20, 2016).
- <sup>72</sup> Virginia Department of Education, "Highlights of Virginia's ESEA Flexibility Request," [http://www.doe.virginia.gov/federal\\_programs/esea/flexibility/highlights.pdf](http://www.doe.virginia.gov/federal_programs/esea/flexibility/highlights.pdf) (accessed March 21, 2016).
- <sup>73</sup> Alaska lowered its n-size from 26 to 5. Arizona lowered its n-size from 40 to 30. Connecticut lowered its n-size from 40 to 20. California's CORE districts lowered their n-size from 100 to 20. Florida lowered its n-size from 30 to 10. Georgia lowered its n-size from 30 to 15. Idaho lowered its n-size from 34 to 25. Illinois lowered its n-size from 45 to 10. Maine lowered its n-size from 20 to 10. Minnesota lowered its n-size from 40 to 10 for reporting purposes and to 20 for accountability purposes. Mississippi lowered its n-size from 30 to 10. Nevada lowered its n-size from 25 to 10. North Carolina lowered its n-size from 40 to 30. Pennsylvania lowered its n-size from 30 to 11. Rhode Island lowered its n-size from 45 to 20. South Carolina lowered its n-size from 40 to 30 and Texas lowered its n-size from 50 to 25.
- <sup>74</sup> See ESSA, section 1111 (e)(1)(B)(iii)(VIII).
- <sup>75</sup> H. S. Bloom and R. Unterman, *Sustained Progress: New Findings About the Effectiveness and Operation of Small Public High Schools of Choice in New York City* (New York, NY: MDRC, 2013); and American Institutes for Research, *Findings from the Study of Deeper Learning: Opportunities and Outcomes* (Washington, DC: Author, 2014).
- <sup>76</sup> Ibid.
- <sup>77</sup> Ibid.
- <sup>78</sup> M. Roderick et al., *Preventable Failure: Improvements in Long Term Outcomes When High Schools Focused on the Ninth Grade Year* (Chicago, IL: The University of Chicago Consortium on Chicago School Research, 2014), <https://ccsr.uchicago.edu/sites/default/files/publications/On-Track%20Validation%20RS.pdf> (accessed March 10, 2015).
- <sup>79</sup> Diplomas Now, "What We Do," <http://diplomasnow.org/about/what-we-do/> (accessed March 21, 2016).
- <sup>80</sup> W. Corrin et al., *Laying Tracks to Graduation: The First Year of Implementing Diplomas Now* (New York, NY: MDRC, 2014).
- <sup>81</sup> K. Mattern, J. Marini, and E. Shaw, *Are AP Students More Likely to Graduate from College on Time?* (Washington, DC: College Board, 2013), <http://research.collegeboard.org/sites/default/files/publications/2014/1/research-report-2013-5-are-ap-students-more-likely-graduate-college.pdf> (accessed September 14, 2015).
- <sup>82</sup> K. D. Mattern, E. J. Shaw, and M. Ewing, *Advanced Placement Exam Participation: Is AP Exam Participation and Performance Related to Choice of College Major?*, (Washington, DC: College Board, 2011), <http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2011-6-ap-participation-performance-major-choice.pdf> (accessed March 21, 2016).
- <sup>83</sup> M. Gordon, E. VanderKamp, and O. Halic, "International Baccalaureate Programmes in Title I Schools in the United States: Accessibility, Participation and University Enrollment," (Washington, DC: International Baccalaureate Organization, 2015), <http://www.ibo.org/globalassets/publications/ib-research/title-1-schools-research.pdf> (accessed September 3, 2015).
- <sup>84</sup> Ibid.
- <sup>85</sup> M. Webb and C. Gerwin, *Early College Expansion: Propelling Students to Postsecondary Success, at a School Near You*. (Washington, DC: Jobs for the Future, 2014), [http://www.jff.org/sites/default/files/publications/materials/Early-College-Expansion\\_031414.pdf](http://www.jff.org/sites/default/files/publications/materials/Early-College-Expansion_031414.pdf) (accessed January 7, 2015).



ALLIANCE FOR  
EXCELLENT EDUCATION

[all4ed.org](http://all4ed.org)

# National Center Brief

## The Importance of Disaggregating Student Data

April 2012

Disaggregating data means breaking down information into smaller subpopulations. For instance, breaking data down into grade level within school aged students, country of origin within racial/ethnic categories, or gender among student populations are all ways of disaggregating data.

Disaggregating student data into subpopulations can help schools and communities plan appropriate programs, decide which evidence-based interventions to select (i.e. have they been evaluated with the target population), use limited resources where they are needed most, and see important trends in behavior and achievement. Collecting and analyzing data can seem intimidating to someone without a strong statistics background, however, many of the tools you need are readily available. This brief provides:

- An overview of the value of disaggregating data
- Common areas of data to disaggregate
- Examples of how disaggregated data has been used
- Limitations of disaggregating data, particularly data describing students

## The Importance of Disaggregating Data

As Safe Schools/Healthy Students sites, you are already collecting important information about students in your district. In addition to the federally required GPRA measures, many states also use the Youth Risk Behaviors Survey (YRBS), in addition to other smaller student information surveys. These data are incredibly valuable; however, much of it is combined, or aggregated, to represent the student population generally. Disaggregating data can show where aggregate data are masking discrepancies. For example, many schools look at student data separated by race/ethnic group. By looking at these data among smaller subpopulations (disaggregating the data), you can see if outcomes vary by subpopulation and if some subpopulations' strong results are masking others' poorer results.

The American Community Survey of 2006, for example, reported a relatively low rate (14 percent) of Asians achieving less than a high school level education. However, disaggregating the data showed discrepancies. Specifically, among Hmong, Laotian, and Cambodian populations, the rates of achieving less than a high school level education were more than double the 14 percent national average: 39 percent (Hmong), 38 percent (Laotian), and 35 percent (Cambodian) (Khan & Ro, 2009). This information could be used for targeted outreach programs and to better inform teachers and other youth-serving providers about which students are at higher risk for lower academic success, information that could easily be missed by only looking at the broader Asian totals. This information

could also be used to inform any needed adaptations to evidence-based programs used with these populations.

Disaggregating data can also valuably inform *program implementation and monitoring*. For example, if student survey results show a gender divide in truancy rates, it might be efficient and useful to have gender specific targeted drop-out prevention and attendance programs. This could ensure that resources are spent where they are needed most.

Disaggregated data can also provide measures of the *effectiveness and equity of a program or ways to view achievement measures*. For example,

- Is there a gender or racial/ethnic outcome difference among students who participate in a particular evidence based intervention?
- Are students in particular grades or with certain teachers performing better, on average, than other grades?
- Are high socio-economic status students overrepresented in accessing and receiving services?

In this way, disaggregated data can confirm perceptions of what is really occurring (i.e. teachers have noticed that ninth grade students consistently perform better on standardized math tests than their eleventh grade counterparts) or debunk stereotypes (i.e. students of lower SES abuse alcohol and other drugs more than their higher SES classmates).

One area where this type of information is commonly used is to show disproportionate minority contact, the number of times a youth is involved with the court system. In fact, the [Office of Juvenile Justice and Delinquency Prevention](#) (OJJDP) uses a specific indicator to collect this information, called the Relative Rate Index (RRI). The RRI compares rates of contact with the juvenile justice and law enforcement systems at various stages among different groups of youth. It can show if there are differences in arrest rates or court sentences, for example, between racial/ethnic groups that are not explained by simple differences in population numbers.

A similar step was taken by the Department of Health and Human Services (HHS) as part of the Affordable Care Act. In the recently released, *HHS Action Plan to Reduce Racial and Ethnic Health Disparities*, a priority was placed on “ensuring that data collection standards for race, ethnicity, sex, primarily language, and disability status are implemented throughout HHS-supported programs, activities, and surveys” (HHS, 2012). Disaggregated data can be used to see if there are meaningful differences by subpopulations in who is accessing mental services and what treatments are successful. This can inform evidence-based programs focusing on mental health as well as documenting a possibly overlooked need for mental health providers.

Disaggregated data can also be used *to advocate for specific policy changes*, to provide evidence for *targeted funding opportunities*, and to look for *patterns over time* and see if similarities or differences within and among subpopulations are emerging. For example, a 1998 Canadian study found that over 90 percent of suicides in First Nation populations were occurring in just 10 percent of First Nation communities in British Columbia (Chandler & Lalonde, 1998). Without disaggregating the data by community, this critical piece of information could have easily been missed. Resources could have been spent too broadly or not focused on the root causes of this discrepancy. Instead, this information

allowed researchers to obtain specific funding to look into what factors were contributing to these substantial population differences. Their results showed that a high level of self-determination was found to significantly reduce a communities' suicide rate (National Collaborating Centre for Aboriginal Health, 2009).

An evaluation specialist is also a valuable resource in this area. They can help to determine what data sets are important, the best way to collect data, and then can assist in analysis and disaggregation.

## Common Areas to Disaggregate

Choosing what data to disaggregate largely depends on the question you are trying to answer about your population and the type of data you have collected. Common characteristics used to disaggregate data include (Boeke, 2012):

- Race/ethnicity (country of origin)
- Generation status (i.e. first, second, etc. generation or recently arrived)
- Immigrant/ refugee status (refugee status often means people are eligible for certain services)
- Age group
- Gender
- Grade
- Geographic (within a state there is often enough data to compare school district data versus a state comparison to a national average)
- Sexual orientation
- Free or reduced lunch status (as a SES indicator)
- Insurance status

## Limitations of Data Disaggregation

Beyond the budgetary and expertise constrictions that many schools now face, there are limitations to what data can be collected, and thus, how data can be analyzed. A big limitation is low statistical power related to small sample sizes when you start disaggregating data. Statisticians from the National Evaluation Team caution that power analyses should be conducted on sample-based data sets, and in the absence of such analyses these data should not be disaggregated further than a cell size of 20 (e.g., if data from a sample size of 70 are disaggregated by race, and there are 20 nonwhites and 50 whites, then that might be okay; but if there are 10 nonwhites and 60 whites, then any conclusions may be misleading. The chance that those 10 nonwhites over-represent a variable of interest compared to the true value of that variable in the nonwhite population is too great). Common limitations to disaggregating data include :

## To protect individual student privacy

Example: A school administers a student survey that collects demographic information on race/ethnicity. The survey items also ask about previous contact with child welfare. If there are two white students in fourth grade and one reported case of previous contact with child welfare by a student who self-reported as non-Hispanic white, it would be easy for someone reviewing those results to identify the student, thus violating the student's privacy.

## Small numbers make it hard to view trends

Example: When evaluating a five-year grant program it would be hard to see true trends when combining three of the five years as a subpopulation. The differences in years could be big enough to misguide what is actually happening by chance or due to program implementation.

## Different data sources do not use the same definitions or break downs

Example: One survey may identify youth by ages 18-24, whereas another would include 18-25 year olds. This could also result from a lack of awareness of visibility of potential significant sub-populations.

## Conclusion

Disaggregating data is important to reveal patterns that can be masked by larger, aggregate data. Looking specifically at sub-populations can help make sure that resources are spent on the areas and students where they are most needed and can have the biggest impact. Perhaps most importantly, disaggregated data can help to make wiser future implementation decisions and secure targeted funding as you work to sustain SS/HS practices.

## Resources

### [Data-Driven High School Reform: The Breaking Ranks Model](#)

Chapter Five: Data-Driven Reform in Low-Performing High Schools

### [Improving School Board Decision-Making: The Data Connection.](#)

Chapter Three: [What is disaggregated data?](#)

Sample Resource Mapping Websites:

[Mapping the Measure of America](#)

[Diversity Data](#)

[Data Resource Center for Child & Adolescent Health](#)

[U.S. Census Interactive Map](#)

[Kids Count Data Center](#)

Safe Schools/Healthy Students Resources by Topic: [Demographic Data](#)

Safe Schools/Healthy Students Resources by Topic: [Cultural and Linguistic Competence](#)

Safe Schools/Healthy Students Resources by Topic: [Evaluation](#)

Safe Schools/Healthy Students Resources by Topic: [Sustainability and Financing](#)

## References:

1. Asian & Pacific Islander American Health Forum. *Disaggregation of Data: Needs of Challenges for Collecting and Reporting Race/Ethnicity Data*. (August 20, 2009). Webinar. Suhaila Khan and Marguerite Ro.
2. United States Department of Health and Human Services. HHS Action Plan to Reduce Racial and Ethnic Health Disparities: A Nation Free of Disparities in Health and Health Care. Washington, 2012. Retrieved from [http://minorityhealth.hhs.gov/npa/files/Plans/HHS/HHS\\_Plan\\_complete.pdf](http://minorityhealth.hhs.gov/npa/files/Plans/HHS/HHS_Plan_complete.pdf).
3. Chandler, M., Lalonde, C. (1998). *Cultural continuity as a hedge against suicide in Canada's First Nations*. *Transcultural Psychiatry*, 35(2): 191-219.
4. National Collaborating Centre for Aboriginal Health. (2009). *The Importance of Disaggregated Data*. Child & Youth Health.
5. Melissa Boeke. Personal communication, January 30, 2012.



DIGITAL SECURITY [[HTTPS://WWW.ACCESSNOW.ORG/ISSUE/DIGITAL-SECURITY/](https://www.accessnow.org/issue/digital-security/)]    PRIVACY [[HTTPS://WWW.ACCESSNOW.ORG/ISSUE/PRIVACY/](https://www.accessnow.org/issue/privacy/)]

# Understanding differential privacy and why it matters for digital rights

25 OCTOBER 2017 | 11:46 AM

“Differential privacy” is a powerful, sophisticated, often misunderstood concept and approach to preserving privacy that, unlike most privacy-preserving tech, **doesn’t rely on encryption**. It’s fraught with complications and subtlety, but it shows great promise as a way to collect and use data while preserving privacy. Differentially private techniques can strip data of their identifying characteristics so that they can’t be used by anyone — hackers, government agencies, and even the company that collects them — to compromise any user’s privacy. That’s important for anyone who cares about protecting the rights of at-risk users, whose **privacy is vital for their safety**. Ideally, differential privacy will enable companies to collect information, while reducing the risk that it will be accessed and used in a way that harms human rights.

Apple is using this approach in its new operating systems, but last month, an [article in Wired](https://www.wired.com/story/apple-differential-privacy-shortcomings/) [<https://www.wired.com/story/apple-differential-privacy-shortcomings/>] exposed ways that the company’s implementation could be irresponsible and potentially insecure. The article is based on the findings of a recent [academic paper](https://arxiv.org/pdf/1709.02753.pdf) [<https://arxiv.org/pdf/1709.02753.pdf>], which delved deep into the MacOS to discover previously undisclosed details about how Apple built its system. The tech behind Apple’s privacy-preserving data collection is fundamentally sound; as usual, the devil is in the details.

In a three-part series, we’ll describe the way private companies, like Apple, both properly and improperly use differential privacy tools. We’ll explain how companies can adopt differential privacy tools responsibly and how lawmakers can respond appropriately.

In this first post, we’ll look at **what differential privacy is and how it works**. In the second, we’ll explore the issues that make it complicated, in Apple’s case and beyond, and [\*\*address some common misunderstandings\*\*](#)



[\[https://www.accessnow.org/differential-privacy-part-2-complicated/\]](https://www.accessnow.org/differential-privacy-part-2-complicated/) . Finally, in part three, we'll look at the **ways differential privacy is used in practice and what responsible use might look like** [\[https://www.accessnow.org/differential-privacy-part-3-extraordinary-claims-require-extraordinary-scrutiny/\]](https://www.accessnow.org/differential-privacy-part-3-extraordinary-claims-require-extraordinary-scrutiny/) .

## What is differential privacy?

Differential privacy isn't, in itself, a technology. It's a *property* that describes some systems — a mathematical guarantee that your privacy won't be violated if your data are used for analysis. A system that is *differentially private* allows analysis while protecting sensitive data behind a veil of uncertainty.

Differential privacy is a way of asking questions about sensitive data. Let's say one party, Alice, has access to private information; another, Bob, wants to know something about it. Alice doesn't want to give Bob access to her customers' data, so instead they make an arrangement: Bob will ask questions ("queries") and Alice will give randomized answers that are *probably* close to the real ones. Bob gets approximations of the answers he wants, but doesn't learn enough to compromise anyone's privacy. Everyone wins.

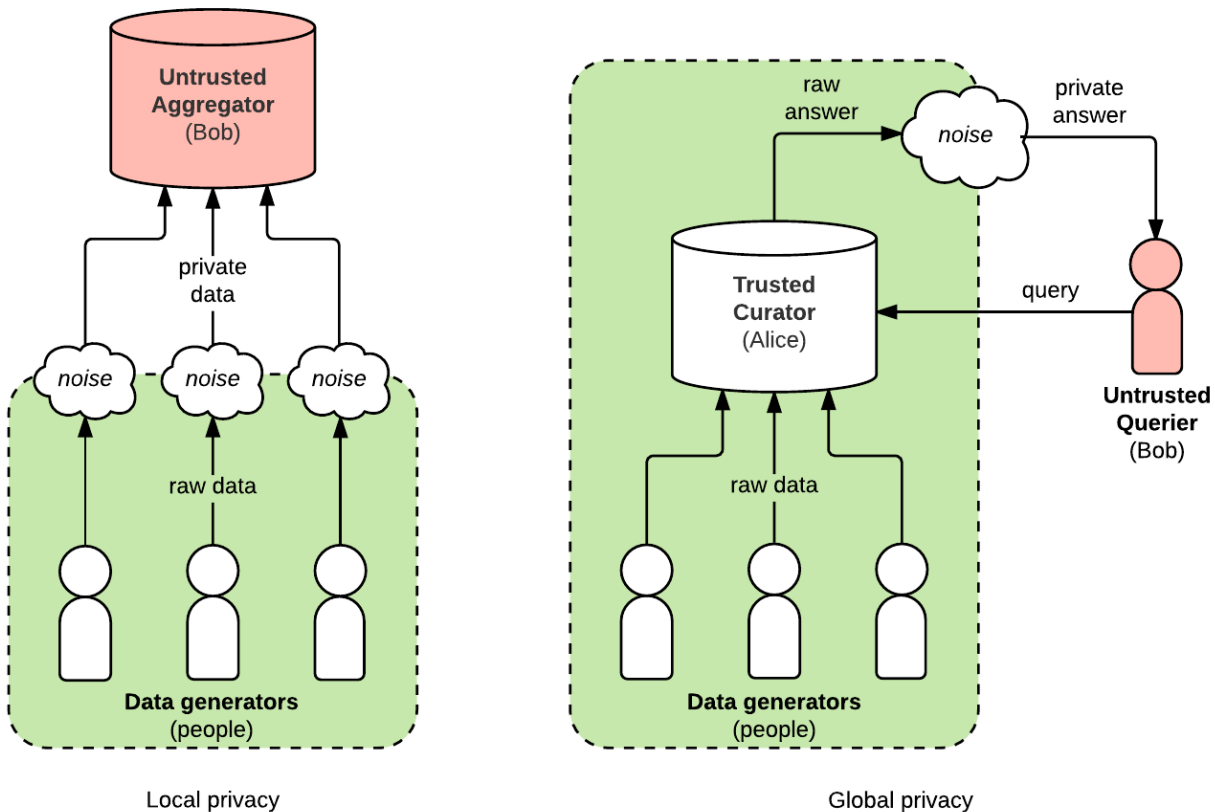
To get technical, differential privacy guarantees that whatever answer Alice might give based on the dataset she has, she would have been almost as likely to give the exact same answer if any one person's data had been excluded. For anyone thinking about giving their data to Alice, this should be encouraging. Whatever Bob asks, Alice's answer is likely to be the same whether that person's data are present or not. It implies that Bob can't use a query to learn much about anyone's data *no matter what else he knows* — even the data of *everyone else* in the database.

## Global vs. local

In general, there are two ways a differentially private system can work: with *global privacy* and with *local privacy*. In a globally private system, one trusted party, whom we'll call the *curator* — like Alice, above — has access to raw, private data from lots of different people. She does analysis on the raw data and adds noise to answers after the fact. For example, suppose Alice is recast as a hospital administrator. Bob, a researcher, wants to know how many patients have the new Human Stigmatization Virus (HSV), a disease whose symptoms include inexplicable social marginalization. Alice uses her records to count the real number. To apply global privacy, she chooses a number at random (using a *probability distribution*, like the [Laplacian](https://en.wikipedia.org/wiki/Laplace_distribution) [\[https://en.wikipedia.org/wiki/Laplace\\_distribution\]](https://en.wikipedia.org/wiki/Laplace_distribution) , which both parties know). She adds the random "noise" to the real count, and tells Bob the noisy sum. The number Bob gets is likely to be very close to the real answer. Still, even if Bob knows the HSV status of all but one of the patients in the hospital, it is mathematically impossible for him to learn whether any particular patient is sick from Alice's answer.

With local privacy, there is no trusted party; each person is responsible for adding noise to their own data before they share it. It's as though each person is a curator in charge of their own private database. Usually, a locally private system involves an untrusted party (let's call them the *aggregator*) who collects data from a big group of people at once. Imagine Bob the researcher has turned his attention to politics. He surveys everyone in his town, asking, "Are you or have you ever been a member of the Communist party?" To protect their privacy, he has each participant flip a coin in secret. If their coin is heads, they tell the truth, if it's tails, they flip again, and let that coin decide their answer for them (heads = yes, tails = no). On average, half of the participants will tell the truth; the other half will give random answers. Each participant can plausibly deny that their response was truthful, so their privacy is protected. Even so, with enough answers, Bob can accurately estimate the portion of his community who support the dictatorship of the proletariat. This technique, known as "random response," is an example of local privacy in action.

Globally private systems are generally more accurate: all the analysis happens on "clean" data, and only a small amount of noise needs to be added at the end of the process. However, for global privacy to work, everyone involved has to trust the curator. Local privacy is a more conservative, safer model. Under local privacy, each individual data point is extremely noisy and not very useful on its own. In very large numbers, though, the noise from the data can be filtered out, and aggregators who collect enough locally private data can do useful analysis on trends in the whole dataset. The diagram below shows the difference between local and global privacy. In both cases, raw data stay safely protected within the green box, and untrusted red parties only see noisy, differentially private information.



[<https://www.accessnow.org/cms/assets/uploads/2017/10/local-vs-global-cropped.png>]

Image credit: Bennett Cyphers

## Epsilon ( $\epsilon$ ): the magic number

Differentially private systems are assessed by a single value, represented by the Greek letter epsilon ( $\epsilon$ ).  $\epsilon$  is a measure of how private, and how noisy, a data release is. Higher values of  $\epsilon$  indicate more accurate, less private answers; low- $\epsilon$  systems give highly random answers that don't let would-be attackers learn much at all. One of differential privacy's great successes is that it reduces the essential trade-off in privacy-preserving data analysis — accuracy vs. privacy — to a single number.

Each differentially private query has an associated  $\epsilon$  that measures its *privacy loss*. Roughly, this measures how much an adversary can learn about anyone's data from a single query. Privacy degrades with repeated queries, and epsilons add up. If Bob make the same private query with  $\epsilon = 1$  twice and receives two different estimates, it's as if he's made a single query with a loss of  $\epsilon = 2$ . This is because he can average the answers together to get a more accurate, less privacy-preserving estimate. Systems can address this with a privacy "budget:" an absolute limit on the privacy loss that any individual or group is allowed to accrue. Private data curators have to be diligent about tracking who queries them and what they ask.

Unfortunately, there's not much consensus about what values of  $\epsilon$  are actually “private enough.” Most experts agree that values between 0 and 1 are very good, values above 10 are not, and values between 1 and 10 are various degrees of “better than nothing.” Furthermore, the parameter  $\epsilon$  is exponential: by one measure, a system with  $\epsilon = 1$  is almost three times more private than  $\epsilon = 2$ , and over 8,000 times more private than  $\epsilon = 10$ . Apple was allegedly using privacy budgets as high as  $\epsilon = 14$  *per day*, with unbounded privacy loss over the long term.

These are the fundamentals for how differential privacy works. Stay tuned for our next post, when we'll dig more deeply into privacy budgets, and talk about a few of the ways differentially private systems can fail to stay private.



**BENNETT CYPHERS**

[<https://www.accessnow.org/profile/bennett-cyphers/>]



2020 accessnow.org

Media Usage



# De-Identification & Student Data

**Reg Leichty**  
Foresight Law + Policy  
**Brenda Leong**  
Future of Privacy Forum

August 2015

## **De-Identification and Student Data**

### **Understanding De-Identification of Education Records and Related Requirements of FERPA**

Appropriate and well-designed student data use by schools, families, researchers, and service providers, greatly enhances teaching and learning. New technologies linked to high capacity broadband networks offer educators and other stakeholders access to powerful analytical tools, rich data, and dynamic digital resources, which can improve student outcomes and inform important education policy reforms. These technology advancements, however, also invite new risks for exposing personally identifiable student data to unauthorized disclosures, misuse, and abuse. In order to reap technology's benefits without encountering these pitfalls, educational agencies and institutions, and their outside partners, must develop and implement more effective strategies and tools for promoting students' privacy and confidentiality.

Data de-identification represents one privacy protection strategy that should be in every student data holder's playbook. Integrated with other robust privacy and security protections, appropriate de-identification – choosing the best de-identification technique based on a given data disclosure purpose and risk level – provides a pathway for protecting student privacy without compromising data's value. This paper provides a high level introduction to: (1) education records de-identification techniques; and (2) explores the Family Educational Rights and Privacy Act's (FERPA) application to de-identified education records.<sup>1</sup> The paper also explores how advances in mathematical and statistical techniques, computational power, and Internet connectivity may be making de-identification of student data more challenging and thus raising potential questions about FERPA's long-standing permissive structure for sharing non-personally identifiable information.

### **The Three-Legged Stool of De-Identification: Personally Identifiable Information, De-identification Strategies, and Data Sharing Purposes & Disclosure Risk Assessment**

Data de-identification is a technically and legally complex issue with special nuances across industries and areas of law. This paper narrowly examines the issue from the perspective of education records and FERPA. The U.S. Department of Education's Privacy and Technical Assistance Center (PTAC) defines de-identification as the "process of removing or obscuring any personally identifiable information from student records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them."<sup>2</sup>

Understanding PTAC's definition is critical to complying with FERPA and ensuring adherence to de-identification best practice. With that goal in mind, this section introduces three core student data de-identification concepts drawn from PTAC's definition and FERPA (law and regulations): personally identifiable information (PII); de-identification processes; disclosure purpose and risk assessment.

---

<sup>1</sup> Family Educational Rights and Privacy Act, 20 U.S.C. 1232g.

<sup>2</sup> *Data De-identification: An Overview of Basic Terms*. U.S. Department of Education Privacy Technical Assistance Center, PTAC-GL, Oct 2012 (updated May 2013).

### *Personally Identifiable Information*

Educational agencies and institutions, and their partners, use de-identification to sever or obscure connections between useful education data and “personally identifiable data.” FERPA’s sharing prohibitions and requirements (explored later in the paper) only apply to PII. In other words, non-personally identifiable information may be shared and retained without restriction (with a narrow exception related to de-identified data connected to a record locator). As a result, understanding the law’s definition of PII is critical to making determinations about how student data may be used, when, and by whom. Under FERPA, PII includes, but is not limited to:

- a) The student’s name
- b) The name of the student’s parent or other family members;
- c) The address of the student or student’s family;
- d) A personal identifier, such as the student’s social security number, student number, or biometric record;
- e) Other indirect identifiers, such as the student’s date of birth, place of birth, and mother’s maiden name;
- f) 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 knowledge of the relevant circumstances, to identify the student with reasonable certainty; or
- g) Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates.<sup>3</sup>

Educational agencies or institutions, and partner entities, such as technology vendors, community based organizations, or researchers, interested in using de-identification as a privacy protection strategy, must pay particular attention to the definition’s inclusion of “indirect identifiers” and “other information.” Data de-identification techniques are used to remove the direct identifiers described above, as well as indirect identifiers and other information, which if left unaddressed, could be used to identify individual students. Other examples of indirect identifiers include race, religion, weight, activities, employment information, medical information, education information, and financial information.<sup>4</sup>

### *Data De-Identification Techniques*

Data de-identification – removing or obscuring PII - begins with eliminating all direct student identifiers from an education record, but education agencies and institutions, and other data holders, must take further steps to ensure that indirect identifiers or other information do not enable an unauthorized actor from determining a student’s identity. These further steps involve using sophisticated mathematical and statistical de-identification techniques, including

---

<sup>3</sup> FERPA, 10 U.S.C. 1232g; 34 CFR § 99.3.

<sup>4</sup> See Privacy and Technical Assistance Online Glossary: <http://ptac.ed.gov/glossary>. Last visited, April 12, 2015.



leveraging technology to ensure the methods are accurately and comprehensively applied across large and complex data sets. Selection of an appropriate de-identification strategy will vary based on specific context, including whether it will be applied to individual level data (information collected and recorded separately for each student) or aggregate data (data combined from several measurements). The former requires much more robust protections.

The U.S. Department of Education's PTAC provides helpful guidance materials, including case studies, that provide detailed information about de-identification approaches,<sup>5</sup> but common methods include the following strategies.<sup>6</sup> See Addendum A for high level examples of each technique.

<b>Blurring</b>	<b>Perturbation</b>	<b>Suppression</b>
Reducing the precision of disclosed data to minimize the certainty of individual identification. For example converting continuous data elements into categorical elements that subsume unique cases.	Making small changes to the data to prevent identification of individuals from unique or rare population groups. For example, swapping data among individual cells to introduce uncertainty.	Removing data, for example from a cell or row, to prevent the identification of individuals in small groups or those with unique characteristics. Usually requires suppression of non-sensitive data.

### *Sharing Purpose & PII Disclosure Risk assessment*

Educational agencies and institutions planning to use de-identification techniques to enable unconsented data sharing – in instances when a FERPA disclosure exception does not apply - must make a “reasonable determination that the student’s identity is not personally identifiable because of unique patterns of information about the student whether through single or multiple releases, and taking into account other reasonably available information.”<sup>7</sup> The standard for making this determination is discussed later in the paper, but neither FERPA, nor the U.S. Department of Education’s FERPA regulations, provide a “safe harbor” listing specific steps that lead to appropriate de-identification. Instead, federal policy provides a standard for making case-by-case judgments of PII disclosure risk at the educational agency, institution, or approved party level.<sup>8</sup> This case-by-case approach means that the list of indirect identifiers that must be removed or obscured to achieve appropriate de-identification will likely vary by circumstance.

---

<sup>5</sup> Privacy and Technical Assistance Center: <http://ptac.ed.gov>. For example, *Frequently Asked Questions on Disclosure Avoidance*, PTAC-FAQ-2, October 2012 (updated May 2013), *Data De-identification: An Overview of Basic Terms*, PTAC-GL, Oct 2012 (updated May 2013), *Case Study #5: Minimizing Access to PII: Best Practices for Access Controls and Disclosure Avoidance Techniques*, PTAC-CS-5, October 2012.

<sup>6</sup> See also, Federal Committee on Statistical Methodology’s Statistical Policy Working Paper 22 Report on Statistical Disclosure Limitation Methodology, (73 Fed. Reg. 74806-35, Dec 9, 2008).

<sup>7</sup> 73 FR 73833, December 9, 2008.

<sup>8</sup> 73 FR 74834, December 9, 2008.



Selecting an appropriate de-identification method depends in part on examining the planned data sharing purpose. The data sharing purpose and de-identification strategy must be compatible.<sup>9</sup> For example, researchers interested in examining students' performance over time might require access to detailed, accurate academic information spanning several years (limiting use of de-identification techniques that diminish a data's validity). Researchers studying a student cohort's growth toward a state's college and career ready standards using a specific pedagogy, for example, would not be able to use data de-identified using a technique that limits the data's reliability and validity. (Alternatively, this type of longitudinal research might be conducted using de-identified data linked to a record locator to enable the originating educational agency or institution to provide de-identified data for the same students over time. Use of such a locator does not render the data "personally identifiable" under FERPA, but it does trigger special requirements.) Conversely, data shared for purposes that require less data precision and accuracy, such as software training or technology research and development, could use much more aggressive de-identification strategies, such as using techniques that replace sensitive information with inauthentic or modified data.

Please note, using de-identification techniques as a privacy tool does not always involve removing all PII, but in situations when PII remains part of a given data set (i.e. where the data has not been completely de-identified), unconsented sharing may only occur with consent or consistent with an appropriate FERPA exception. For example, an educational agency or institution sharing PII under a qualified FERPA exception may wish to use de-identification techniques to minimize PII released to an outside entity, even though they may lawfully share a range of student level information. To be more specific, a researcher might conduct a study that requires a discrete list of indirect identifiers that together could lead to the student's identification, such as a student's age, race and family financial information, but not requiring other PII found in the same education records. In such an instance, these three pieces of personally identifiable student data – and other information attached them - would remain subject to FERPA's disclosure limitations and other requirements, but de-identification techniques (e.g., suppression) could provide additional protection for the student by removing data, for example from a cell or row, unnecessary to the study. Researchers lawfully using PII in this context and other cases, however, must completely de-identify any report or other information before releasing it to the public or other parties, including other researchers.<sup>10</sup>

Entities planning to use de-identification techniques must mitigate the risk of exposing the identity of individual students. Therefore, after examining the requirements of a given data sharing purpose, education data holders must also assess the risks associated with their planned disclosure, including considering past data releases (the risk of re-identification is cumulative), sample size, the nature of the data recipient,<sup>11</sup> whether the data will be further shared or made

---

<sup>9</sup> *Data De-identification: An Overview of Basic Terms*. U.S. Department of Education Privacy Technical Assistance Center, PTAC-GL, Oct 2012 (updated May 2013), p. 4.

<sup>10</sup> 73 FR 74834, December 9, 2008.

<sup>11</sup> The Department of Education has said "there is no statutory authority in FERPA to modify the prohibition on disclosure of personally identifiable information from education records, or the exceptions to the written consent requirement, based on the track record of the party, including journalists and researchers, in maintaining the confidentiality of information from education

public, and other contextual conditions.<sup>12</sup> More aggressive de-identification strategies are required in situations when the student data is potentially at greater risk of re-identification.

For example, de-identified data shared for a specific purpose with a trusted public or private entity such as a state department of education, institution of higher education, or professional vendor with strict legal and contract protections (e.g., an agreement with strict re-disclosure limitations), might be less likely to be widely available later (decreasing the re-identification threat associated with cumulative data releases), compared for example to annual school or district performance data posted directly to a public website to comply with federal and state accountability requirements. Why is greater public availability of a properly de-identified data set a potential problem? In some cases, de-identified data might be subject to nefarious comparisons with other data sets (e.g., with widely available student “directory information”) or other attempts to reveal PII. When data enters the public domain, it could be exposed to cutting-edge tools and techniques designed to compare the de-identified data to other publicly available data sets and thus reveal a student’s identity (the FERPA implications of such a breakthrough are discussed further below).

Although experts disagree about the extent to which new technologies and techniques can “back map” de-identified data to reveal a student’s identity, a serious statistical analysis that ensures all direct and indirect identifiers have been removed can be performed to ensure any re-identification risk is remote.

In short, prudent student data holders should consider using – in light of new data mining and comparison techniques that might be more effective than is commonly accepted – the most aggressive de-identification strategies possible when data will be made public or shared widely. When data is shared with limited restricted parties under strong controls and under a FERPA exception, a combination of technical, administrative and contractual controls will be appropriate for reasonable de-identification measures that may preserve greater utility of the data.

### **Application of FERPA to De-Identified Records**

As a general rule, FERPA prohibits the disclosure of education records containing personally identifiable student data without parent or eligible student consent.<sup>13</sup> Therefore, the release of education records that have been appropriately de-identified – purged of direct and all necessary indirect identifiers in a given context - is not considered a “disclosure” under FERPA, since by definition such records do not contain PII.<sup>14</sup> Properly de-identified student data thus may be shared without limitation under FERPA (although other federal and state privacy laws may apply). Furthermore, “de-identified information from education records is not subject to any

---

records that they have received.” (73 FR 74834). Nonetheless, the recipients’ identity should likely be considered among other variables in each risk assessment.

<sup>12</sup> *Frequently Asked Questions – Disclosure Avoidance*, p. 4, PTAC-FAQ-2, Oct 2012 (updated May 2013). p.2-3

<sup>13</sup> 20 U.S.C. 1232g(b)(1)

<sup>14</sup> 34 CFR 99.31(b)(1)

destruction requirements because, by definition, it is not ‘personally identifiable information.’<sup>15</sup> The Department has said, however, a party releasing de-identified student data might mitigate risks associated with future data releases by independently requiring data destruction in some circumstances.<sup>16</sup>

There is one important exception, however, to FERPA’s unconsented sharing exception for de-identified data. De-identified data coupled with a record code or locator by an educational agency or institution – allowing it to be matched later to the record source - may only be shared for education research. Although the Department’s regulations and guidance do not specifically discuss the question, it appears that educational agencies or institutions may select any qualified third party to conduct research under this provision, but all secondary (non-research) uses of de-identified data with a record locator are prohibited. Furthermore, the data sharing entity may not disclose information about how it generated and assigned the record code, or other information that might allow a data recipient to identify a student based on the record code. Lastly, the record code must not be based on a student’s social security number or other personal information.<sup>17</sup> Such a data set remains categorized as “de-identified,” and may thus be shared without parent or eligible student consent, but unlike other de-identified data it may only be shared for the research purpose specified to the educational agency or institution, consistent with the other requirements described above.

Before such data sharing can occur, however, the education record must be properly de-identified. As referenced above, the “releasing party is responsible for conducting its own analysis and identifying the best methods to protect the confidentiality of information from education records it chooses to release.”<sup>18</sup> This determination depends on FERPA’s disclosure risk assessment standard. This standard asks whether a “reasonable person in the school community who does not have personal knowledge of the relevant circumstances” could use the released data, and other publicly available data, to identify an individual student with “reasonable certainty.”<sup>19</sup> This standard extends to possible data holders beyond the literal school community.

The Department of Education does not require educational agencies and institutions to use specific data disclosure avoidance techniques to achieve this standard, and stated in a recent rulemaking, “it is not possible to prescribe or identify a single method to minimize the risk of disclosing personally identifiable information that will apply in every circumstance...”<sup>20</sup> The Department has also said “determining whether a particular set of methods for de-identifying data and limiting disclosure risk is adequate cannot be made without examining the underlying data sets, other data that have been released, publicly available directories and other data that are linked or linkable to the information in questions.”<sup>21</sup> In other words, the party releasing data

---

<sup>15</sup> 73 FR 15585, March 24, 2008

<sup>16</sup> 73 FR 74835, December 9, 2008

<sup>17</sup> 34 CFR 99.31(b)(2)(i)-(iii).

<sup>18</sup> 73 FR 74835, December 9, 2008.

<sup>19</sup> 34 CFR § 99.3, 34 CFR §99.31(b)(1)

<sup>20</sup> 73 FR 74835, December 9, 2008

<sup>21</sup> Ibid at 74835

must perform a context specific analysis and identify the best method for protecting student information subject to disclosures. Proper application of the accepted mathematical and statistical de-identification strategies described earlier in the paper meet this legal standard in many instances, but by law each sharing context must be independently analyzed against the Department's reasonableness standard.<sup>22</sup>

Some experts have argued that given recent cases where researchers have leveraged access to other publicly available data sets to identify specific individuals, absolute data de-identification may be impossible, or at a minimum, increasingly difficult.<sup>23</sup> In light of this uncertainty, data sharing parties should very carefully analyze each proposed disclosure of de-identified data against FERPA's reasonableness standard and also consider using contracts that specify protections – above and beyond FERPA – that could further minimize the risk of re-identification.

### **De-Identified Data: Retention and Destruction**

FERPA permits third party data holders, including vendors, to retain and use appropriately de-identified data – so long as it is not associated with a record locator -for any secondary purpose. Furthermore, FERPA does not describe how de-identified data should be managed, including, as described above, when and how the data should be destroyed. Vendors and other third party holders must, however, ensure that a given de-identified data set is not subject to relevant contract terms, or other Federal, state, and local privacy laws and regulations, which might contain more stringent data retention or destruction requirements.<sup>24</sup> For example, personal data subject to the Children's Online Privacy Protection Act may only be retained so long as is necessary to fulfill the purpose for which it was collected, and COPPA covered entities must delete the information using reasonable measures to protect against its unauthorized access or use.<sup>25</sup>

Although FERPA does not govern the use, retention and destruction of properly de-identified data, third parties should have sound policies – guided by National Institute of Standards and Technology or PTAC best practice recommendations - addressing these issues. This internal, independent step includes ensuring that de-identified data is destroyed when it is no longer needed, in order to minimize re-identification risks associated with possible future efforts to compare and link the data with other data sets. Data holders must also ensure that they take proper actions to destroy data. Simply deleting data is not sufficient in most cases and PTAC's data destruction best practices provide helpful guidance. PTAC recommends that data holders “make risk-based decisions on which [destruction] method - [e.g. clearing, purging, or destroying data] - is most appropriate based on the data type, risk of disclosure, and the impact if that data were to be disclosed without authorization.”<sup>26</sup> The data de-identification method used to remove

---

<sup>22</sup> 34 CFR 99.31.(b)(1). See also, PTAC *Frequently Asked Questions – Disclosure Avoidance*, p. 4, PTAC-FAQ-2, Oct 2012 (updated May 2013).

<sup>23</sup> *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, Paul Ohm, University of Colorado Law School, [UCLA Law Review](#), Vol. 57, p. 1701, 2010 .

<sup>24</sup> Privacy and Technical Assistance Center, *Best Practices for Data Destruction*, p. 5, PTAC-IB-5, May 2014.

<sup>25</sup> 16 C.F.R. § 312.10.

<sup>26</sup> PTAC Best Practices for Data Destruction, p. 5.

PII from a data set should be a central factor in making this determination. Data holders seeking additional guidance on proper destruction strategies should consult recommendations made by the National Institute of Standards and Technology and other expert sources.<sup>27</sup>

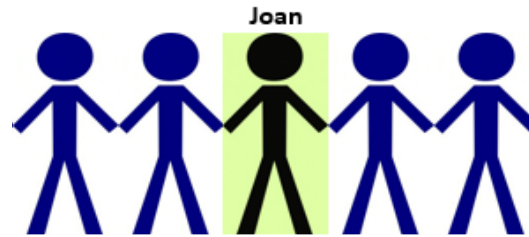
## **Conclusion**

De-identification offers an important tool for educational agencies, institutions and their partners seeking to maximize student data's potential value to improving teaching and learning, while also carefully protecting student privacy and confidentiality. Proper data de-identification requires, however, deep technical knowledge and expertise and adherence to industry best practice. Therefore, student data holders should not attempt to de-identify student data sets without competent support. They should also consult competent legal counsel to ensure that their data management policies and practices – including de-identification strategies - comply with FERPA and all other relevant federal, state, and local laws and requirements potentially applicable to the data they manage.

---

<sup>27</sup> National Institute of Standards and Technology (NIST) Special Publication 800-88 Rev. 1: Guidelines for Media Sanitization. December 2014.

## Illustration of Common De-Identification Measures in Aggregate Data Sets



Raw Individual Student Data in Aggregate Data Table

<b>Joan's Director Identifiers</b> Student Name: Joan Smith Students Parents: John Smith & Jackie Smith Address: 0000 00 <sup>th</sup> Street, Washington,D.C. Student Number: 4444 Social Security Number: 555-555-555	<b>Joan's Indirect Identifiers</b> Data of Birth: 11/01/2000 Race: Alaska Native Gender: Female Place of Birth: Washington, D.C. Family Income: \$85,000 GPA: 3.75
--	--

Redacted Individual Student Level Data in Aggregate Data Table

<b>All Direct Identifiers Removed</b>	<b>Joan's Indirect Identifiers</b> Data of Birth: 11/01/2000 Race: Alaska Native Gender: Female Place of Birth: Washington, D.C. Family Income: \$85,000 GPA: 3.75
---------------------------------------	--

Blurring (Reducing Data Precision including Using Broader Categories)

<b>All Direct Identifiers Removed</b>	<b>Joan's Indirect Identifiers</b> Data of Birth: 2000 Race: Minority Gender: Female Mother's Maiden Name: Johnson Place of Birth: Mid-Atlantic Family Income: \$50,000 - \$100,000 GPA: 3.5 – 4.0
---------------------------------------	---

Perturbation (Small Data Changes, including through Swapping Data among Cells)

<b>Mike's Indirect Identifiers</b> Data of Birth: 1999 Race: Unique Characteristic Removed Gender: Female Mother's Maiden Name: Unique Characteristic Removed Place of Birth: Midwest Family Income: \$50,000 - \$100,000 GPA: 3.5 – 4.0	<b>Joan's Indirect Identifiers</b> Data of Birth: 2000 Race: Unique Characteristic Removed Gender: Male Mother's Maiden Name: Unique Characteristic Removed Place of Birth: Northeast Family Income: \$50,000 - \$100,000 GPA: 3.5 – 4.0
---	---

Suppression (Removing Data from a Cell or Row)

<b>All Direct Identifiers Removed</b>	<b>Joan's Indirect Identifiers</b> Data of Birth: 2000 Race: Unique Characteristic Removed Gender: Female Mother's Maiden Name: Unique Characteristic Removed Place of Birth: Mid-Atlantic Family Income: \$50,000 - \$100,000 GPA: 3.5 – 4.0
---------------------------------------	--