Frequently Asked Questions—Disclosure Avoidance

Overview

The U.S. Department of Education established the Privacy Technical Assistance Center (PTAC) as a "one-stop" resource for education stakeholders to learn about data privacy, confidentiality, and security practices related to student-level longitudinal data systems. PTAC provides timely information and updated guidance on privacy, confidentiality, and security practices through a variety of resources, including training materials and opportunities to receive direct assistance with privacy, security, and confidentiality of longitudinal data systems. More PTAC information is available on http://ptac.ed.gov.

Purpose

This document is intended to provide general guidance to State and local educational agencies and institutions about the best practice strategies for protecting personally identifiable information from education records (PII) in aggregate reports. The paper provides suggestions on how to ensure that necessary confidentiality requirements are met, including compliance with the Family Educational Rights and Privacy Act (FERPA). The information is presented in the form of responses to frequently asked questions (FAQs), followed by a list of additional resources at the end.

Please note that the current brief document is designed to highlight key issues surrounding the use of disclosure avoidance methods. The U.S. Department of Education plans to conduct additional training on best practices for data disclosure avoidance, which will cover specific strategies in greater depth.

FAQs: Disclosure Avoidance of Personally Identifiable Information in Aggregate Reporting

Question: *What is the definition of "disclosure" and "disclosure avoidance"*?

Answer: "Disclosure" means to permit access to or the release, transfer, or other communication of PII by any means. Disclosure can be authorized, such as when a parent or an eligible student gives written consent to share education records with an authorized party (e.g., a researcher). Disclosure can also be unauthorized or accidental. An unauthorized disclosure can happen due to a data breach or a loss (see PTAC's Data Security: Top Threats to Data Protection brief at http://ptac.ed.gov/sites/default/files/issue-brief-threats-to-your-data.pdf for more information and security tips). An accidental disclosure can occur when data released in public aggregate reports are unintentionally presented in a manner that allows individual students to be identified.

"Disclosure avoidance" refers to the efforts made to reduce the risk of disclosure, such as applying statistical methods to protect PII in aggregate data tables. These safeguards, often referred to as

disclosure avoidance methods, can take many forms (e.g., data suppression, rounding, recoding, etc.).

Question: *If I am only publishing aggregate data tables, do I still need to be concerned about disclosure avoidance?*

Answer: Yes. The aggregation of student-level data into school-level (or higher) reports removes much of the risk of disclosure, since no direct identifiers (such as a name, Social Security Number, or student ID) are present in the aggregated tables. Some risk of disclosure does remain, however, in circumstances where one or more students possess a unique or uncommon characteristic (or a combination of characteristics) that would allow them to be identified in the data table (this commonly occurs with small ethnic subgroup populations), or where some easily observable characteristic corresponds to an unrelated category in the data table (e.g., if a school reports that 100% of males in grade 11 scored at "Below Proficient" on an assessment). In these cases, some level of disclosure avoidance is necessary to prevent disclosure in the aggregate data table.

Question: *What legal obligation do educational agencies and institutions have to protect PII in aggregate reports?*

Answer: Under FERPA, educational agencies and institutions reporting or releasing data derived from education records are responsible for protecting PII in the reports from disclosure. The U.S. Department of Education also states, in reporting achievement results under section 1111(h) of the Elementary and Secondary Education Act of 1965, as amended (ESEA), to "not use disaggregated data for one or more subgroups… to report achievement results… if the results would reveal personally identifiable information about an individual student" and to "implement appropriate strategies to protect the privacy of individual students" (34 CFR §200.7). Further, "to determine whether disaggregated results would reveal personally identifiable information about an individual student" (34 CFR §200.7), States are instructed to follow FERPA requirements (34 CFR §99).

Question: *What issues should educational agencies and institutions consider to successfully balance privacy protection requirements with data disclosure requirements?*

Answer: Since the release of any data carries at least some element of risk, it may not possible to entirely eliminate the risk of accidental data disclosure. However, organizations disclosing the data in the form of public aggregate reports are responsible for minimizing any such risk while still meeting the disclosure requirements and providing as much useful and transparent information to the public as possible. Before each planned release of student data, an organization must determine the acceptable level of risk of disclosure. This means that in each specific case, the entity disclosing the data should evaluate the risk of PII disclosure within the context that the data will be used, and

choose a safeguard strategy that is the most appropriate for that particular context.

Question: *Is public reporting of data for small groups ("small cells") the same thing as a disclosure?*

Answer: Reporting unrounded frequency counts in small cells, such as an exact number of students in a small group, does not by itself constitute a disclosure; however, the smaller the cell size, the greater the likelihood that someone might be able to identify an individual within that cell, and thus the greater the risk of disclosure. Many statisticians consider a cell size of 3 to be the absolute minimum needed to prevent disclosure, though larger minimums (e.g., 5 or 10) may be used to further mitigate disclosure risk.

Question: *What standard is used to evaluate disclosure risk?*

Answer: The FERPA standard for de-identification assesses whether a "reasonable person in the school community who does not have personal knowledge of the relevant circumstances" could identify individual students based on reasonably available information, including other public information released by an agency, such as a report presenting detailed data in tables with small size cells (34 CFR §99.3 and §99.31(b)(1)). The "reasonable person" standard should be used by State and local educational agencies and institutions to determine whether statistical information or records have been sufficiently redacted prior to release such that a "reasonable person" (i.e., a hypothetical, rational, prudent, average individual) in the school community should not be able to identify a student because of some well-publicized event, communications, or other similar factor. School officials, including teachers, administrators, coaches, and volunteers, are not considered in making the reasonable person determination since they are presumed to have inside knowledge of the relevant circumstances and of the identity of the students.

Question: *What are some of the commonly used disclosure avoidance techniques?*

Answer: Some of the most commonly used disclosure avoidance methods include data suppression, blurring, and perturbation. When deciding which method to apply in a specific situation, it is important to evaluate the different methods in terms of their effects on the utility of the data and the risk of disclosure.

- *Suppression* 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. This method may often result in very little data being produced for small populations, and it usually requires additional suppression of non-sensitive data to ensure adequate protection of PII (e.g., complementary suppression of one or more non-sensitive cells in a table so that the values of the suppressed cells may not be calculated by subtracting the reported values from the row and column totals). Correct application of this technique generally results in low risk

of disclosure; however, it can be difficult to perform properly because of the necessary calculations (especially for large multi-dimensional tables). Further, if additional information related to the suppressed data is available elsewhere, the suppressed cells may potentially be re-calculated.

- *Blurring* is used to reduce the precision of the disclosed data to minimize the certainty of identification. Examples of blurring include rounding, aggregating across different populations or geographies, and reporting percentages and ranges instead of exact counts. This method may affect the utility of the data by reducing users' ability to make inferences about small changes in the data. Similarly, blurring methods that rely on aggregation across geographies or subgroups may interfere with time-series or cross-sectional data analysis. Applying this technique generally ensures low risk of disclosure; however, if any unblurred cell counts or row and/or column totals are published (or are available elsewhere), it may be possible to calculate the values of sensitive cells.

- *Perturbation* involves making small changes to the data to prevent identification of individuals from unique or rare population groups. Examples of this technique include swapping data among individual cells (this still preserves the marginal distributions, such as row totals) and introducing "noise," or errors (e.g., by randomly reclassifying values of a categorical variable). This method helps to minimize the loss of data utility as compared to other methods (e.g., compared to the complete loss of information due to suppression); however, it also reduces the transparency and credibility of the data. Therefore, perturbation is often considered inappropriate for public reporting of program data, from an accountability perspective. Applying this technique generally ensures low risk of disclosure, as long as the rules used to alter the data (e.g., the swapping rate) are protected. This requires securing the information about the technique itself as well as restricting access to the original data, so that perturbation rules cannot be reverse-engineered.

Question: *Does the U.S. Department of Education require educational agencies and institutions to use specific data disclosure avoidance techniques?*

Answer: The Department does not mandate a particular method, nor does it establish a particular threshold for what constitutes sufficient disclosure avoidance. These decisions are left up to the individual State and local educational agencies and institutions to determine what works best within their specific contexts.

As a general recommendation, in aggregate publically available reports, whenever possible, data about individual students (e.g., proficiency rates presented as cross-tabulated tables) 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.

Moreover, under the ESEA, each State must establish a minimum sub-group size (e.g., number of

students in a table cell) below which it will not publically report assessment data. This threshold value and other reporting rules should be specified in the documents describing the State's data reporting policies and practices implemented to protect student privacy, such as in the State Accountability Workbook (www.ed.gov/admins/lead/account/stateplans03/index.html). Minimum cell sizes adopted by the States range from 5 to 30 students, with a majority of States using 10 as their minimum (NCES 2011-603). Please note that simple suppression of small subgroups may not be sufficient to protect the privacy of all students, since the suppressed numbers can often be easily calculated by subtracting the reported subgroups' totals from the all-student totals or by comparing the school and district enrollment information. In some cases, complementary suppression of additional non-sensitive cells may be necessary.

Question: *What practical suggestions can the U.S. Department of Education provide to educational agencies and institutions to help them implement recommended disclosure avoidance techniques?*

Answer: The Department strongly suggests using a computer program to apply disclosure limitation methods, as some techniques may be difficult to implement accurately by hand. In particular, to ensure correct application of data suppression method, care should be taken when suppressing any complementary cells. Lastly, it is preferable, from a data user perspective, to apply consistent methods year to year and to use the same disclosure avoidance strategies for similar types of data releases.

Question: *Does the U.S. Department of Education intend to release more in-depth guidance on data disclosure avoidance techniques in the future? What topics will it cover?*

Answer: Yes, the Department is currently working on developing best practices for States to consider when designing and adopting their own disclosure avoidance strategies. The best practices document will review different disclosure avoidance techniques and their applicability across different contexts, and will be supplemented by examples and definitions of any relevant statistical terminology.

Additional Resources

The resources below include links to federal regulations and several guidance and best practices resources. These include some draft recommendations developed by the National Center for Education Statistics (NCES) in published Technical Briefs. While these recommendations may not be appropriate for every situation, they may provide a better understanding of the issues involved in selecting and applying disclosure avoidance methods to education data.

- *Case Study #5: Minimizing Access to PII: Best Practices for Access Controls and Disclosure Avoidance Techniques.* Privacy Technical Assistance Center (Oct 2012): http://ptac.ed.gov/sites/default/files/case-study5-minimizing-PII-access.pdf

- Code of Federal Regulations - Title 34: Education. *Disaggregation of data*. 34 CFR §200.7: www.gpo.gov/fdsys/pkg/CFR-2011-title34-vol1/pdf/CFR-2011-title34-vol1-sec200-7.pdf

- FERPA regulations, U.S. Department of Education: www.ed.gov/policy/gen/reg/ferpa

- *FERPA regulations amendment*.U.S. Department of Education (December 9, 2008): www.ed.gov/legislation/FedRegister/finrule/2008-4/120908a.pdf

- *FERPA regulations amendment*. U.S. Department of Education (December 2, 2011): www.gpo.gov/fdsys/pkg/FR-2011-12-02/pdf/2011-30683.pdf

- *Frequently Asked Questions—Disclosure Avoidance.* Privacy Technical Assistance Center (Oct 2012): http://ptac.ed.gov/sites/default/files/FAQs_disclosure_avoidance.pdf

- Privacy Technical Assistance Center (PTAC), U.S. Department of Education: http://ptac.ed.gov

- *SLDS Technical Brief 3: Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting* (NCES 2011-603): http://nces.ed.gov/pubs2011/2011603.pdf

- *Statistical Policy Working Paper 22 - Report on Statistical Disclosure Limitation Methodology.* Federal Committee on Statistical Methodology, Office of Management and Budget (1994): http://fcsm.gov/working-papers/wp22.html

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

Glossary

Education Program is defined as any program principally engaged in the provision of education, including, but not limited to, early childhood education, elementary and secondary education, postsecondary education, special education, job training, career and technical education, and adult education, and any program that is administered by an educational agency or institution. For more information, see the Family Educational Rights and Privacy Act regulations, 34 CFR §99.3.

Education records means records directly related to a student and maintained by an educational agency or institution, or by a party acting on behalf of the agency or institution. For more information, see the Family Educational Rights and Privacy Act regulations, 34 CFR §99.3.

Personally identifiable information (PII) from education records includes information, such as a student's name or identification number, that can be used to distinguish or trace an individual's identity either directly or indirectly through linkages with other information. See Family Educational Rights and Privacy Act regulations, 34 CFR §99.3, for a complete definition of PII specific to education records and for examples of other data elements that are defined to constitute PII.