# An Overview of Data Practices for Protecting Personally Identifiable Information (PII)

## Background

The Family Educational Rights and Privacy Act (FERPA) affords parents and all students certain rights with respect to education records. These rights extend to students in all grade levels, beginning in preschool and extending through their years in postsecondary institutions.

As more and more information management systems become electronic and accessible online, it is important for school district officials to be mindful about protecting student privacy and with whom and how student information is shared.

Details regarding the legal aspects of data collection and data sharing can be found on the Kansas State Department of Education web page under the Data Central tab: https://datacentral.ksde.org/

The Shawnee Mission School District has the authority and responsibility to protect student data privacy through de-identification and anonymization of data. Requests for data under the Kansas Open Records Act (KORA) or through other avenues, such as partnerships that involve academic research and data sharing agreements with vendors, must be limited in scope the extent possible that it protects students from incidental identification.

The FERPA 2008 regulations subsection on de-identified records allows for the nonconsensual release of student level information from education records, provided that (1) all personally identifiable information is removed and (2) there is a reasonable determination that a student's identity is not personally identifiable. In making this determination, both single and multiple data releases from the education records should be taken into account along with other information available from other sources (34 CFR § 99.31(b)(1)).

To ensure the confidentiality of student records and information, District staff with access to student data should not disclose or share any aggregate or individual student data other than with District staff who are on a need-to-know basis. Further, District staff should direct all requests for student data, whether formal or informal, to the District's Chief Communication Officer.

## Types of Disclosure

Data reporting practices are in place to avoid any of the three categories of disclosure, which may compromise student data privacy.

**Identity disclosure** occurs if a third party can identify a subject or respondent from the released data. The level or risk of identifying whether an individual is a respondent or subject of a data collection depends on the type of data involved. For example, providing a list of the members of a varsity soccer team discloses student identities, but it does not violate data privacy laws. However, providing a list of students suspended or expelled from school discloses student identities, AND it violates student data privacy.

**Attribute disclosure** occurs when confidential information about a data subject is revealed and can be attributed to the subject. Attribute disclosure occurs when confidential information about a person is revealed **or may be closely estimated**. Thus, attribute disclosure compromises identification of the subject and divulging confidential information about the subject. For example, if a school has a small number of ethnic minority students, and a report indicates that nearly all the students in that minority group performed a certain way on an assessment, a third party may be able to infer to a high degree how each individual's assessment performance.

**Inferential disclosure** occurs when individual information can be inferred with high confidence from statistical properties of the released data. For example, data may show a high correlation between income and home prices. Home prices are typically public information. A third party might use this information to infer the income of a family.

The following pages describe techniques used in the Shawnee Mission School District to limit the chances that a students' information could be revealed. The specific technique used depends on the risk associated with the data attributes. For example, when reporting aggregate data on test scores or discipline infractions, district staff must act conservatively to avoid identity disclosure. In contrast, a demographic breakdown of students on a sports team would not necessarily need to follow the same practices.

Source: https://www.hhs.gov/sites/default/files/spwp22.pdf

## General Practices on Group Size and Group Types

The first step in protecting PII in reports is to establish minimum thresholds for displaying a population or subgroup. However, additional steps may be required to protect PII as indicated in the examples that follow.

### Groups of 30 or Smaller

Following the general practices used by the US Department of Education and the Kansas Department of Education, total group sizes smaller than 30 *may* be redacted or replaced with a non-precise range. This is particularly important when there is additional breakout reporting of this group. An example of "non-precise" may appear as "<30" if working with a count of students or "<10%" or "NA" if working with percentages or other metrics.

### Groups of 10 or Smaller

Following the general practices used by the US Department of Education and the Kansas Department of Education, total and sub-group sizes smaller than 10 *must* be redacted or replaced with a non-precise range. For example, results for group sizes smaller than ten are not displayed on the Kansas Building Report Cards (https://ksreportcard.ksde.org)

### Groups of Size Zero

District staff may need to suppress group sizes of zero. This will primarily occur in instances where the other reported groups are relatively small and where revealing a zero could result in attribute or inferential disclosure.

### Combining Groups

When reporting aggregate data, district staff may choose to combine reasonably similar groups so that the size becomes large enough to be reportable. For example, in many schools the numbers of students who qualify for free lunch or reduced lunch are too small to be reported separately. It is common practice to combine them into a single group in order to make the aggregate data large enough to avoid unintended disclosures.

### De-Identifying Data Sets

In the event that a third party is provided access to a raw data set in which the third party does not have access to PII, that data set is stripped of identifying characteristics such as student name, student ID, and birthdate. The data is also analyzed to evaluate the risk of attribute or incidental disclosure. For example, if an attribute or set of attributes are associated with ten or fewer students, either those records are removed, or the attribute is merged with another attribute.

## Data Display Examples

### Case #1

Table A: Students with One or More Discipline Referrals

| Total | Black | White | Hispanic | Other |
|---|---|---|---|---|
| 32 | 10 | 15 | 2 | 5 |

The sample data provided in Table A leaves the school at risk for disclosure of PII. An individual reading this report may happen to know of two Hispanic students referred to the office. In that case, it reveals that those were the only two, and no other Hispanic students were referred. The data provided here is sharable with the building administrator, who has the legal authority to access PII. This table should not be shared with persons who are not authorized to view PII.

Table B: Students with One or More Discipline Referrals

| Total | Black | White | Hispanic | Other |
|---|---|---|---|---|
| 32 | 10 | 15 | NA | 5 |

The modification offered in Table B shows that simply suppressing the smallest value(s) does nothing to protect PII. Readers can simply subtract from the total students to obtain the missing number. Hence, other techniques are required to avoid PII disclosure.

Table C: Students with One or More Discipline Referrals

| Total | Black | White | Hispanic | Other |
|---|---|---|---|---|
| 32 | 10 | 15 | <10 | <10 |

Table C contains a preferred modification, but it may not be enough to protect unintended disclosure of PII. Readers can deduce that seven students who received one or more referrals were either Hispanic or some other racial category. Depending on the total demographic make-up of the students in this building, this still might not be enough modification to protect PII.

Table D: Students with One or More Discipline Referrals

| Total | Black | White | Hispanic | Other |
|---|---|---|---|---|
| ~30 (approx.) | 10 | 15 | <10 | <10 |

The modifications in Table D are perhaps the most ideal for this scenario in public reporting. The reader has an approximate total group size, which offers some insight into the proportion of students referred to the office by race/ethnicity. However, it prevents readers from being able to deduce the exact number of students in the suppressed fields.

## Case #2

Table A: Kansas Assessment Results in English Language Arts

| | N | Level 1 | Level 2 | Level 3 | Level 4 |
|---|---|---|---|---|---|
| Total | 32 | 4 | 10 | 11 | 7 |
| No Lunch Support | 22 | 0 | 5 | 10 | 7 |
| Free/Reduced Lunch | 10 | 4 | 5 | 1 | 0 |

Table A contains multiple opportunities for PII disclosure. For instance, by reporting that only one Free/Reduced Lunch child scored at Level 3 or higher, anyone who is able to identify that child also knows that every other child in the Free/Reduced group scored below that level. Similarly, if a reader happened to know a child who scored at Level 1, that reader could also deduce that the child was receiving Free/Reduced Lunch support.

Table B: Kansas Assessment Results in English Language Arts

| | N | Level 1 | Level 2 | Level 3 | Level 4 |
|---|---|---|---|---|---|
| Total | 32 | <10 | 10 | 11 | <10 |
| No Lunch Support | 22 | <10 | <10 | 10 | <10 |
| Free/Reduced Lunch | 10 | <10 | <10 | <10 | <10 |

Table B shows one common method for suppressing small group sizes to protect PII. However, strictly hiding all values below ten also suppresses any possible interpretation of student performance.

Table C: Kansas Assessment Results in English Language Arts

| | N | Level 1 | Level 2 | Level 3 | Level 4 |
|---|---|---|---|---|---|
| Total | 32 | <20% | 20-40% | 20-40% | 20-40% |
| No Lunch Support | 22 | <20% | 20-40% | 40-60% | 20-40% |
| Free/Reduced Lunch | 10 | 20-40% | 40-60% | <20% | <20% |

Table C shows one possible compromise using a technique called blurring.  Precise values are replaced with ranges so that readers can draw some conclusions about performance by student group, but risk of PII disclosure remains low.

## Resources

Andreou, A., Goga, O., & Loiseau, P. (2017, July). Identity vs. Attribute Disclosure Risks for Users with Multiple Social Profiles. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 163-170). ACM. https://lig-membres.imag.fr/gogao/papers/information_disclosure_ASONAM2017.pdf

Data Quality Campaign (June 2017). Understanding Minimum N-Size and Student Data Privacy: A Guide for Advocates. https://dataqualitycampaign.org/wp-content/uploads/2017/06/DQC-N-size-paper-FINAL.pdf

Federal Committee on Statistical Methodology. STATISTICAL POLICY WORKING PAPER 22 (Second version, 2005): Report on Statistical Disclosure Limitation Methodology. December 2005. https://www.hhs.gov/sites/default/files/spwp22.pdf.

Kansas State Department of Education (KSDE). Student Data Collection and Security: https://www.ksde.org/Data-Central/Student-Data-Collection-and-Security-in-Kansas.

Seastrom, M. (2017). Best Practices for Determining Subgroup Size in Accountability Systems While Protecting Personally Identifiable Student Information. Institute of Education Sciences Congressionally Mandated Report. IES 2017-147. *Institute of Education Sciences*. https://eric.ed.gov/?id=ED572044

Sullivan, Colleen M. (1992) U.S. Census Bureau. An Overview of Disclosure Principles. https://www.census.gov/srd/papers/pdf/rr92-09.pdf