Privacy Technical Assistance Center

DaSy The Center for IDEA Early Childhood Data Systems

Data De-identification: An Overview of Basic Terms

IDEA Early Childhood Data De-identification Glossary

Overview

The U.S. Department of Education (Department) established two national technical assistance centers, the Privacy Technical Assistance Center (PTAC) and the Center for IDEA Early Childhood Data Systems (DaSy), to assist States in complying with the privacy, security, and confidentiality requirements of the Family Educational Rights and Privacy Act (FERPA) and the Individuals with Disabilities Education Act (IDEA) as they implement longitudinal data systems. DaSy collaborated with PTAC to adapt and develop resources to help IDEA Part C early intervention and Part B 619 preschool special education staffs address data confidentiality, data sharing, and data privacy questions and issues as they build and use early childhood data systems. This document is part of a series of documents that address data systems issues to specifically meet the needs of IDEA Part C early intervention and Part B 619 preschool special education. Some of the documents in this series were originally developed by PTAC and were adapted by DaSy for this audience.

Purpose

Data de-identification is important to ensure that States protect the confidentiality rights of children with disabilities and their families as States share data that include personally identifiable information to improve service delivery, child outcomes, program accountability, and related purposes. This document is intended to assist early intervention service programs and providers and preschool special education programs and agencies in maintaining compliance with privacy and confidentiality requirements under IDEA and FERPA. It reviews the terminology used to describe data de-identification as well as related concepts and approaches. In addition to defining and clarifying the distinctions among several key terms, the document provides general best practice de-identification strategies for different types of data and statistical techniques that can be used to protect children against data disclosures. Additional resources on applicable IDEA and FERPA requirements are also identified.

Data De-identification—Key Concepts and Strategies

Data de-identification is important because two statutes administered by the U.S. Department of Education, FERPA and IDEA require the privacy of personally identifiable information (PII) in the education records of students and in the education and early intervention records[1] of children with disabilities. These statutes also generally require that parental consent be obtained before the disclosure of such information. However, if such information is properly "de-identified," it may be shared publicly. To avoid unauthorized disclosure of PII from education and early intervention records in public reporting, children's data must be adequately protected at all times. For example, when IDEA Part C early intervention service programs and providers and Part B 619 preschool special education programs and agencies at the local or State level publicly report on child outcomes on an aggregated basis, they must apply disclosure avoidance strategies to prevent unauthorized release of information about individual children. For successful data protection, it is essential that techniques are appropriate for the intended purpose and that their application follows the best practices.

Thus, a vital step is to evaluate disclosure limitation techniques against the desired level of data protection. To aid educational agencies, programs, and institutions with making these decisions and to help ensure consistency of the terminology used by the education and early childhood community, PTAC and DaSy provide this alphabetized list of techniques commonly used to protect privacy of individual child

[1] Note that under IDEA Part C regulation §303.414(b)(2), an *early intervention record* is equivalent to an *education record* under FERPA. For clarity, we use *education and early intervention records* in the rest of this document.

education and early intervention records and the types of redacted data files that can be produced by applying them. Figure 1 provides an overview of the main types of data typically managed and disseminated by educational and early intervention organizations, programs, and agencies by level of sensitivity and associated need for protection.

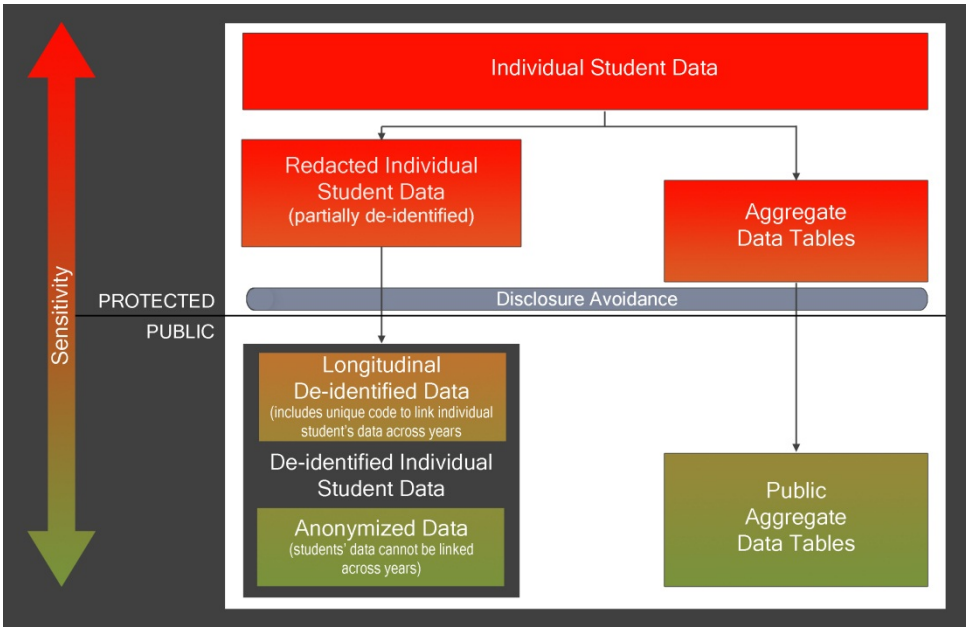

Figure 1: Types of data by sensitivity and need for protection from unauthorized or inadvertent disclosure. Source: PTAC, "Data De-identification: An Overview of Basic Terms."

Anonymization [of data] refers to the process of data de-identification which produces de-identified data. If data are properly de-identified, an individual reviewing the data would not be able to identify individual children. As such, anonymized data are not useful for monitoring the progress of individual children or a small group of children (i.e. specific disability, race/ethnicity); however, they can be used for other research or professional development purposes. It is important to review the resulting file to ensure that additional disclosure limitation methods do not need to be applied. The documentation for the anonymized data file should identify any disclosure limitation techniques that were applied and their implications for the analysis.

De-identification [of data] refers to the process of removing or obscuring any personally identifiable information from children's education and early intervention records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them. Specific steps and methods used to de-identify information (see "Disclosure limitation method" for details) may vary depending on the circumstances, but they should be appropriate to protect the confidentiality of the individuals. While it may not be possible to remove the disclosure risk completely, de-identification is considered successful when there is no reasonable basis to believe that the remaining information in the education and early intervention records can be used to identify an individual.

De-identified data may be shared without the consent required by FERPA (34 CFR §99.30) with any party for any purpose, including parents, the general public, and researchers [34 CFR §99.31(b)(1)], and thus are also an exception under IDEA Part B [34 CFR §300.622(a)] and Part C [34 CFR §303.414(b)]. These data are typically released in the form of aggregated data (such as tables showing numbers of children receiving services by race, age, and sex, which are typically collected and reported by State and local programs under IDEA section 618) or as microdata (such as individual-level child outcome data collected under IDEA sections 616 and 642 by IDEA Part C early intervention and Part B 619 preschool special education programs). Individual-level data may be released with or without an attached education and early intervention record code (the education and early intervention record code cannot be based on the children's personal information), which allows education researchers to track performance of individual children without the children's identity being revealed to

them [34 CFR §99.31(b)(2)]. Researchers can use the code only to match individual education and early intervention records across previously de-identified data files from the same source (e.g., to compare child outcomes from the same IDEA Part C early intervention program over several years); they cannot use the code to access the original data source without parental consent. (Note that coded individual-level data can be released only for the purposes of education research and are subject to certain conditions—see education and early intervention "Record code" for more information.) De-identified data that do not include an education or early intervention record code and cannot be linked to the original data source are referred to herein as *anonymized*.

It is important to note that PII may include not only direct identifiers, such as names, child IDs or Social Security numbers, but also any other sensitive and non-sensitive information that, alone or combined with other information that is linked or linkable to a specific individual, would allow identification. Therefore, simple removal of direct identifiers from the data to be released *does not* constitute adequate de-identification. Properly performed de-identification involves removing or obscuring all identifiable information until all data that can lead to individual identification have been expunged or masked. The standard under the FERPA regulations is that a reasonable person in the school community (or in the case of the IDEA, the preschool or early intervention community) who does not have personal knowledge of the relevant circumstances could not identify an individual child based on the information released.

Further, when making a determination about whether the data have been sufficiently de-identified, it is necessary to consider cumulative re-identification risk from all previous data releases and other reasonably available information, including publicly available directory information and de-identified data releases from education and early intervention records as well as other sources. In particular, care should be taken to monitor new releases of de-identified individual-level child data that are released with an attached education or early intervention record code.

De-identification strategy. See "Disclosure limitation method."

Disclosure means to permit access to or the release, transfer, or other communication of PII by any means (34 CFR §99.3). Disclosure can be authorized, such as when a parent gives written consent to share education and early intervention records with an authorized party (e.g., a researcher). Disclosure also can be unauthorized or inadvertent (accidental). An unauthorized disclosure can happen because of 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 children to be identified.

It is important to note that the release of education and early intervention records that have been de-identified is not considered a "disclosure" under FERPA (and IDEA), since by definition, de-identified data do not contain PII that can lead to identification of individual children. This statement holds true regardless of whether de-identified data have been released with an attached education or early intervention record code or without it; however, releases of coded de-identified data are subject to certain conditions (see "Record code" for more information).

Disclosure avoidance refers to the efforts made to de-identify data in order to reduce the risk of disclosure of PII. The choice of the appropriate de-identification strategy (also referred to as disclosure limitation method) depends on the nature of the data release, the level of protection offered by a specific method, and the usefulness of the resulting data product. The two major types of data release are aggregated data (such as tables showing numbers of children receiving services by race, age, and sex, which data collection and reporting occurs under IDEA section 618) and microdata (such as individual-level child outcome data collected under IDEA sections 616 and 642 by an IDEA Part C early intervention or Part B 619 preschool special education program). Several acceptable de-identification methods exist for each type of data (see "Disclosure limitation method" for details).

Disclosure limitation method (also known as disclosure avoidance method) is a general term referring to a statistical technique used to manipulate data before release to minimize the risk of inadvertent or unauthorized disclosure of PII. Entities releasing data should apply a consistent de-identification strategy to all their data releases of a similar type (e.g., tabular and individual-level data) and similar sensitivity level. Organizations should document their data reporting rules in the documents describing their data reporting policies and privacy protection practices, such as a *Data Governance Manual*. (See

PTAC's "Data Governance and Stewardship" brief at http://ptac.ed.gov/sites/default/files/issue-brief-data-governance-and-stewardship.pdf for more information on best practices in data governance.)

The major methods used by the Department for disclosure avoidance for tabular data include defining a minimum cell size (meaning no results will be released for any cell of a table with a number smaller than "X" or else cells are aggregated until no cells with a number smaller than "X" remain) and controlled rounding (meaning that cells with a number smaller than "X" require that numbers in the affected rows and columns be rounded so that the totals remain unchanged). Whenever possible, data about individual children (e.g., proficiency scores) are combined with data from a sufficient number of other children to disguise the attributes of a single child. When this is not possible, data about small numbers of children are suppressed.

For releases of child-level data, the primary consideration is whether the proposed release contains any individuals with unique characteristics whose identity could be deduced by the combination of variables in the file. If such a condition exists, one of a number of methods is used. These include data blurring, such as "top-coding" a variable (e.g., family income above a certain level is recorded to a defined maximum) and applying various data perturbation techniques.

For additional guidance on specific steps and acceptable methods for de-identifying data on children, see the list of resources at the end of the document.

The following are examples of disclosure limitation methods:[2]

- **Blurring**—Used to reduce the precision of the disclosed data to minimize the certainty of individual identification. There are many ways to implement blurring, such as by converting continuous data elements into categorical data elements (e.g., creating categories that subsume unique cases), aggregating data across small groups of respondents, and reporting rounded values and ranges instead of exact counts to reduce the certainty of identification. Another approach involves replacing an individual's actual reported value with the average group value; this may be performed on more than one variable with different groupings for each variable. This method may reduce the user's ability to make inferences about small changes in the data. It also reduces the user's ability to perform time-series or cross-case analysis. Correct application of this technique generally ensures low risk of disclosure; however, if row/column totals are published (or available elsewhere), someone might be able to calculate the actual values of sensitive cells.
- **Masking**—Used to "mask" the original values in a data set to achieve data privacy protection. This general approach uses various techniques, such as data perturbation, to replace sensitive information with realistic but inauthentic data or modify original data values based on predetermined masking rules (e.g., by applying a transformation algorithm). The purpose of masking is to retain the structure and functional usability of the data while concealing information that could lead to the identification, either directly or indirectly, of an individual child. Masked data are used to protect individual privacy in public reports and can be a useful alternative when the real data are not required, such as in user training or software demonstration. Specific masking rules may vary depending on the sensitivity of the data and organizational data disclosure policies.
- **Perturbation**—Involves making small changes to the data to prevent identification of individuals from unique or rare population groups. Data perturbation is a data masking technique in that it is used to mask the original values in a data set to avoid disclosure. Examples of this statistical technique include swapping data among individual cells to introduce uncertainty, so that the data user does not know whether the real data values correspond to certain education and early intervention records, and introducing "noise," or errors (e.g., by randomly misclassifying values of a categorical variable). This method can minimize loss of utility compared with other methods. However, it is deemed inappropriate for program data because it reduces the transparency and credibility of the data, which can have enforcement and regulatory implications. Also, a person with access to some (e.g., a single State's) original data may be able to reverse-engineer the perturbation rules used to alter the data.

[2] PTAC *Protection of Personally Identifiable Information through Disclosure Avoidance Techniques*, February 16, 2012: http://ptac.ed.gov/sites/default/files/discl-avoid-pres-feb10-12_0.pdf

- **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. Suppression may 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 suppression generally ensures low risk of disclosure; however, it can be difficult to perform properly because of the necessary calculations (especially for large multidimensional tables). Further, if additional data are available elsewhere (e.g., total child counts are reported), the suppressed data may be recalculated.

Record code is the unique descriptor used to match individual-level education and early intervention records across de-identified data files from the same source (e.g., for the purposes of examining outcomes of individual children over time). Under FERPA [34 CFR §99.31(b)(2)], an education agency, institution, or party that has received education and early intervention records or information from those records, may release de-identified child-level data (microdata) for education research purposes by attaching a code to each education and early intervention record that may allow the researcher to match information received from the same source under the specified conditions. The coded de-identified microdata are to be used only for education research purposes, the party receiving the data is not allowed any access to information about how the descriptor was generated and assigned, and the record code cannot be used to identify the children or to match the information from education and early intervention records with data from any other source. Furthermore, a record code may not be based on a child's Social Security number or other personal information.

Redaction is the process of expunging sensitive data from the education and early intervention records before disclosure in a way that meets established disclosure requirements applicable to the specific data disclosure occurrence (e.g., removing or obscuring PII from published reports to meet Federal, State, and local privacy laws as well as organizational data disclosure policies). (See “Disclosure limitation method” for more information about specific techniques that can be used for data redaction.)

Additional Resources

The Department 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 their website: http://ptac.ed.gov.

The Center for IDEA Early Childhood Data Systems (DaSy) is a national technical assistance center funded by the Department’s Office of Special Education Programs (OSEP). DaSy works with States to support IDEA Part C early intervention and Part B 619 preschool special education State programs in the development or enhancement of integrated early childhood longitudinal data systems. DaSy’s work is organized around three areas to support IDEA Part C early intervention and Part B 619 preschool special education State staff: (1) generating new knowledge and useful products for States to use in the development and enhancement of statewide early childhood data systems; (2) designing and implementing a continuum of technical assistance strategies with States that are evidence based, relevant, useful, and cost-effective; and (3) providing national leadership and coordination on early childhood data systems. More information about DaSy is available on their website http://dasycenter.org/.

Please direct questions to the PTAC at mailto:PrivacyTA@ed.gov or 855-249-3072 and/or DaSy at mailto:dasycenter@sri.com or 650-859-3881.

The resources below include links to Federal regulations and several guidance documents providing in-depth descriptions of techniques that can be used to de-identify tabular as well as child-level data. 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 relevant concepts and issues involved in selecting and applying data de-identification methods to education data.

- "Best Practices for Access Controls and Disclosure Avoidance Techniques Webinar." Privacy Technical Assistance Center (Nov 2012): http://ptac.ed.gov/sites/default/files/Webinar_DD_Nov2012Final.pdf
- "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
- Federal regulations resources, U.S. Department of Education: www.ed.gov/policy/gen/reg/edpicks.jhtml?src=ln
- 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
- Guidance on the Amendments to the Family Educational Rights and Privacy Act by the Uninterrupted Scholars Act: http://www2.ed.gov/policy/gen/guid/fpco/ferpa/uninterrupted-scholars-act-guidance.pdf
- IDEA Part B regulations, U.S. Department of Education (2006): http://www.ecfr.gov/cgi-bin/text-idx?SID=d74c644d5aeea44a16267317b21601be&node=34:2.1.1.1.1&rgn=div5
- IDEA Part C regulations, U.S. Department of Education (2011): http://www.ecfr.gov/cgi-bin/text-idx?SID=d74c644d5aeea44a16267317b21601be&node=34:2.1.1.1.2&rgn=div5
- "Letter to Edmunds (December 2012)." U.S. Department of Education's Office of Special Education Programs response regarding whether or not "early intervention records" under IDEA Part C are considered "education records" under FERPA: http://www2.ed.gov/policy/speced/guid/idea/memosdcltrs/edmunds.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
- *SLDS Technical Brief 1: Basic Concepts and Definitions for Privacy and Confidentiality in Student Education Records* (NCES 2011-601): http://nces.ed.gov/pubs2011/2011601.pdf
- *SLDS Technical Brief 3: Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting* (NCES 2011-603): http://nces.ed.gov/pubs2011/2011603.pdf

Glossary

The 2014 IDEA Part C regulations clarified the relationship between terms used under IDEA Part C and terms used under FERPA. Below is the translation of terms as clarified under IDEA Part C regulation 34 CFR §303.414(b)(2).

For a side-by-side comparison of the primary legal provisions and definitions in IDEA Part B, IDEA Part C, and FERPA that relate to the requirement to protect the confidentiality of personally identifiable information of students and children served under the IDEA, please see the IDEA and FERPA Confidentiality Provisions crosswalk available: http://www2.ed.gov/policy/gen/guid/ptac/pdf/idea-ferpa.pdf.

Crosswalk of Terms

FERPA	IDEA Part C
education record	early intervention record
education	early intervention
educational agency or institution	participating agency
school official	qualified early intervention service (EIS) personnel/service coordinator
State educational authority	lead agency
student	child under IDEA Part C

FERPA Definition

- **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.
- **Early childhood education program** means- (a) A Head Start program or an Early Head Start program carried out under the Head Start Act (42 U.S.C. 9831 *et seq.*), including a migrant or seasonal Head Start program, an Indian Head Start program, or a Head Start program or an Early Head Start program that also receives State funding; (b) A State licensed or regulated child care program; or (c) A program that—(1) Serves children from birth through age six that addresses the children's cognitive (including language, early literacy, and early mathematics), social, emotional, and physical development; and (2) Is—(i) A State prekindergarten program; (ii) A program authorized under section 619 or Part C of the Individuals with Disabilities Education Act; or (iii) A program operated by a local educational agency. For more information, see the Family Educational Rights and Privacy Act regulations, 34 CFR §99.3.
- **Educational agency or institution** means any public or private agency or institution to which this part applies under §99.1(a). (Authority: 20 U.S.C. 1232g(a)(3)). For more information, see the Family Educational Rights and Privacy Act regulations, 34 CFR §99.3.

IDEA Part B and Part C Definitions

- **Child**, as defined by Part C regulations, means an individual under the age of six and may include an infant or toddler with a disability, as that term is defined in 34 CFR §303.21. For more information, see the Individual with Disabilities Education Act regulations, 34 CFR §303.6.
- **Child with a disability**, as defined by Part B regulations, means a child having mental retardation, a hearing impairment (including deafness), a speech or language impairment, a visual impairment (including blindness), a serious emotional disturbance (referred to in this part as "emotional disturbance"), an orthopedic impairment, autism, traumatic brain injury, another health impairment, a specific learning disability, deaf-blindness, or multiple disabilities, and who, by reason thereof, needs

special education and related services. For more information, see the Individual with Disabilities Education Act regulations, 34 CFR §300.8.

- **Education records**, as defined by Part B regulations, mean the type of records covered under the definition of "education records" in 34 CFR part 99 (the regulations implementing the Family Educational Rights and Privacy Act of 1974, 20 U.S.C. 1232g [FERPA]). (34 CFR §300.611)
- **Early intervention records**, as defined by Part C regulations, mean all records regarding a child that are required to be collected, maintained, or used under Part C of the Act [IDEA] and the regulations in this part. (34 CFR §303.403)
- **Participating agency**, as defined by Part B regulations, means any agency or institution that collects, maintains, or uses personally identifiable information, or from which information is obtained, under Part B of the Act [IDEA]. (34 CFR §300.611)
- **Participating agency**, as defined by Part C regulations, means any individual, agency, entity, or institution that collects, maintains, or uses personally identifiable information to implement the requirements in Part C of the Act [IDEA] and the regulations in this part with respect to a particular child. A participating agency includes the lead agency and EIS [early intervention service] providers and any individual or entity that provides any Part C services (including service coordination, evaluations and assessments, and other Part C services), but does not include primary referral sources, or public agencies (such as the State Medicaid or CHIP [Children's Health Insurance Program]) or private entities (such as private insurance companies) that act solely as funding sources for Part C services. (34 CFR §303.403)
- **Personally identifiable**, as defined by Part B regulations, means information that contains: (a) the name of the child, the child's parent, or other family member; (b) the address of the child; (c) a personal identifier, such as the child's social security number or student number; or (d) a list of personal characteristics or other information that would make it possible to identify the child with reasonable certainty. (34 CFR §300.32)
- **Personally identifiable information**, as defined by Part C regulations, means personally identifiable information as defined in 34 CFR §99.3 [See FERPA], as amended, except that the term "student" in the definition of personally identifiable information in 34 CFR §99.3 means "child" as used in this part and any reference to "school" means "EIS [early intervention service] provider" as used in this part. (34 CFR §303.29)

The contents of this document were developed under a grant from the U.S. Department of Education, #H373Z120002 and in collaboration with the following offices and centers:

- At the U.S. Department of Education:
  - Office of Special Education and Rehabilitative Services (OSERS), Office of Special Education Programs (OSEP)
  - OSERS' Office of Policy and Planning (OPP)
  - Office of the General Counsel (OGC)
  - Office of Management (OM), Privacy, Information, and Records Management Services (PIRMS)
  - OM's Family Policy Compliance Office (FPCO)
- The Center for IDEA Early Childhood Data Systems (DaSy)
- The Privacy Technical Assistance Center (PTAC)

Although many offices within the U.S. Department of Education provided input into, and review of, the content in this document to make available technical assistance on best practices, the content should not be read as representing the policy of, or endorsement by, the U.S. Department of Education. For further information, you may contact the DaSy grant project officers, Meredith Miceli and Richelle Davis.