ISSUE BRIEF U.S. Policy

Privacy Enhancing Technologies for Education Researchers

May 2026

FUTURE OF PRIVACY FORUM

AUTHOR

Jim Siegl, Senior Fellow, FPF

About This Document

This document serves as a guide for researchers and education agencies on utilizing Privacy Enhancing Technologies (PETs) to conduct rigorous academic studies while safeguarding student identities. It outlines the specific utility and limitations of tools like synthetic data, differential privacy, and secure multiparty computation, highlighting the necessary "privacy-accuracy tradeoff" inherent in each method. The summary concludes that while PETs provide mathematically verifiable ways to minimize data exposure, they must be integrated into strong human governance frameworks to ensure both research integrity and student protection.

Primary audiences: Education Researchers and Policy Makers.

The Future of Privacy Forum (FPF) is a non-profit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. Learn more about FPF by visiting fpf.org.

Privacy Enhancing Technologies (PETs) for Researchers

Executive Summary

Educational research answers questions about program effectiveness, equity, access, and long-term outcomes. Researchers typically need student-level data and enough context to ensure their research is effective and accomplishes their goal. At the same time, disclosing student data to researchers may increase the risk of reidentification. The risk is not limited to direct identifiers such as names or student IDs; uncommon or rare characteristics and small groups can also reveal too much.

Some analyses—particularly those that rely on predefined queries or aggregate statistics—can be completed without direct access to student-level data. Agencies can enable analysis through protected environments or Privacy-Enhancing Technologies (PETs). PETs are methods that help reduce the risks of sharing student data while also preserving its value. They can lower the amount of sensitive data shared, support safer publication of results, and protect small groups. However, PETs do not replace good governance. It's necessary to implement data minimization, role-based access controls, auditability, and enforceable agreements.

How PETs Can Help Researchers

PETs can be applied to reduce the amount of sensitive data shared or disclosed**.** Instead of sending full datasets, PETs can allow computation where the data already resides or within a secure enclave. PETs can also allow you to publish findings with less risk of identifying students**.** Techniques such as differential privacy can add "noise" to outputs, making it harder to reidentify students with rare or unique characteristics.

Protecting small groups when reporting is important to ensure data stays protected and cannot be linked back to any specific student. When subgroup sizes are small, PETs can reduce the chance that tables, dashboards, or any other outputs inadvertently reveal outcomes for a single student. It's also necessary to protect sensitive attributes**.** PETs can help minimize exposure of highly sensitive fields by supporting computations that do not require those attributes to be accessible outside a controlled environment.

Deciding What PETs to Use

The best PET choice depends on the study. A simple report has different needs than a longitudinal analysis or an academic publication. Four practical factors can help guide decisions regarding which PET to use:

- **Research question:** What outcome matters? What level of precision is required? Which outputs will be published?
- **Data sensitivity:** Consider both direct and indirect identifiers, the sensitivity of attributes, and the risk of linkage across systems.
- **Operational need**: Identify privacy risks that need to be mitigated.
- **Privacy-accuracy tradeoff**: Every PET imposes some cost on analytical precision. The acceptable level of that cost depends on whether results will inform decisions requiring high precision, or whether trend-level findings are sufficient. Identifying that threshold early shapes which PET is appropriate and how it should be configured.

The Privacy-Accuracy Tradeoff in Research Contexts

No PET is without some tradeoff. Each approach reduces privacy risk by limiting what researchers can see, compute, or release — and each of those limitations has an analytical cost. That cost takes different forms depending on the method: Differential privacy (DP) introduces statistical noise that degrades precision for small groups; synthetic data may not faithfully reproduce rare characteristics or tail distributions; SMPC and TEEs preserve accuracy within the computation but restrict the types of analyses that can be run and introduce operational complexity that limits iteration.

For researchers, the relevant question is not whether a tradeoff exists, but whether the tradeoff is acceptable given the study's inferential goals. A longitudinal equity analysis examining chronic absenteeism trends across large districts may tolerate meaningful noise. A study attempting to detect small subgroup outcome differences may not. Neither outcome is wrong — but each requires a different configuration or a different PET.

This also has implications for replication and publication. Results produced under DP carry an epsilon parameter that should be reported. Synthetic data findings should disclose generation methodology and validation approach. Reviewers and policymakers reading research outputs need enough information to assess what the privacy protections cost analytically.

Synthetic Data

Synthetic data is artificially generated data that **mimics the statistical properties** of real-world datasets without containing any records from actual individuals. By preserving the **distributions, correlations, and patterns** in the original data, synthetic sets enable robust testing and analysis while significantly mitigating the risk of re-identification.

Synthetic extracts are particularly useful when teams need realistic test data but do not need real identities or exact counts for production decisions. For example, when an SLDS team collaborates with a vendor to build a new reporting tool, rather than sharing extracts, the agency can provide a synthetic dataset that mirrors their **data architecture and logic** without exposing protected student information.

However, synthetic data does not automatically provide formal privacy guarantees, and its analytical validity depends on the generation approach. Synthetic data is generally unsuitable for studies that depend on rare subgroup representation, precise effect sizes, or tail-distribution behavior. A synthetic dataset that accurately reflects aggregate patterns may systematically misrepresent the experiences of small groups — precisely the populations that equity-focused education research often prioritizes. Validation against held-out real data, with documented fidelity metrics, should be a precondition for using synthetic data in published research.

Differential Privacy (DP)

Differential privacy (DP) is a formal privacy framework that bounds how much the inclusion of any one student's data can influence reported results. It does this by introducing carefully calibrated statistical noise, so that outputs are similar whether or not a particular student's record is included, while still preserving overall patterns in the data. The strength of this protection is governed by a parameter (often called ε, or "epsilon") that reflects a tradeoff between privacy and precision.

Researchers should specify an epsilon value before analysis begins, not after. Post-hoc calibration to achieve a desired result undermines both the privacy guarantee and the integrity of the finding. In practice, there is no universal standard for epsilon; the right value depends on dataset size, query sensitivity, and how much precision the research question requires.

DP is particularly useful when publishing dashboards or research results that disaggregate outcomes by subgroup or school. For example, in a statewide report of chronic absenteeism by grade, race/ethnicity, and district, small districts or subgroups can make results sensitive to individual students. Applying DP to reported rates and counts limits the extent to which any one student's status can be inferred from those outputs, while allowing analysts to study broader trends.

Secure Multiparty Computation

Secure Multiparty Computation (SMPC) is a cryptographic approach that enables multiple organizations (such as state agencies) to jointly compute statistics or models over their combined data without sharing raw, student-level records. Each party splits its data into encrypted "shares" and distributes them across participating servers. Computation is performed directly on these shares, and only the agreed-upon results are revealed.

SMPC produces mathematically exact results for agreed-upon computations, so it does not introduce the same noise-based accuracy tradeoff as DP. The tradeoff is instead operational: the analysis must be fully specified before computation runs, and iterative or exploratory analysis is not well-suited to SMPC protocols. Researchers accustomed to interactive data exploration will find this constraining.

Because SMPC protocols are designed to be mathematically equivalent to operations on the underlying data, they can produce exact results for a specified analysis (e.g., sums, averages, regressions) without exposing the inputs during computation. This protects data in use, but it does not, by itself, prevent sensitive information from being inferred from the outputs.

As a result, SMPC is typically combined with governance and disclosure controls. Participating organizations still need clear agreements defining the purpose of the analysis, who is authorized to run computations, and what results can be released, along with appropriate disclosure avoidance rules applied to outputs.

Conclusion

Researchers do not have to choose between research data fidelity and student privacy. With the right approach, both goals can be achieved simultaneously because PETs offer mathematically verifiable tools for conducting studies that minimize data exposure without compromising statistical validity. In practice, this means selecting approaches that align with the research question, the data's sensitivity, and the acceptable level of precision—recognizing that different PETs introduce trade-offs among utility, complexity, and risk.

At the same time, PETs do not eliminate the need for disclosure review or governance. They reduce exposure at different stages of the research lifecycle, but privacy risk ultimately depends on what can be inferred from released outputs. It also depends on what was lost analytically—and responsible research practice requires documenting both. Clear rules for access, computation, and publication remain necessary, particularly for small groups, linked datasets, and iterative analyses.

For education agencies and researchers, the implication is practical: PETs are most effective when integrated into existing data governance frameworks rather than treated as standalone solutions. Used this way, they can enable broader collaboration, support more flexible research designs, and strengthen confidence that student data is being used in ways that are both analytically rigorous and appropriately protective.

Washington, DC | Brussels | Singapore

FPF.org