

Education Research Center (ERC) Supplemental Data Submission & Processing Procedure

Purpose

This procedure outlines the steps for ERC researchers submitting supplemental data (publicly available, aggregate, or individual-level) for merging with ERC datasets after Advisory Board approval.

For this procedure, **merge with ERC data** means combining any external dataset with ERC-held data at any level (individual, program, school, district, region, or year) in a way that allows analysis across datasets.

Advisory Board Data Dictionary Requirement

A complete Data Dictionary (using the THECB template) is **required** for all supplemental datasets intended for merging, even if the data is publicly available or aggregate.

- The Data Dictionary **must** be submitted as part of the proposal for Advisory Board review.
- Without an approved Data Dictionary, the dataset will **not** enter the processing queue.
- Advisory Board approval does **not** guarantee approval of every data element. Each element will undergo a separate compliance review.

Submission Preparation

After receiving Advisory Board approval, researchers must prepare:

1. A complete **Data Dictionary** (using THECB template) that lists only the variables approved in the project proposal.
2. A signed **Supplemental Data Usage Approval Form**.

Important Notes

- Do **not** submit variables that are redundant with existing ERC datasets (CBM, TEA, NSC, or TWC).
- Submitting variables that fall into or are similar to the Personally Identifiable Information (PII) categories in Appendix A and B will significantly increase review time. These variables require extensive legal review to ensure FERPA compliance and mitigate re-identification risks.

Review and Approval Steps

All inquiries must go through the ERC director to maintain a clear and consistent communication channel.

1. Submit the Data Dictionary and Approval Form to the ERC director.

2. ERC director forwards to THECB; THECB and TEA review each variable for FERPA and legal compliance.
3. Clarification requests are routed back through the ERC director.
4. Final approved Data Dictionary is sent to the researcher; researchers must formally accept it before sending any data files.
5. Dataset must **exactly match** the approved Data Dictionary (variable names, formats, and values).

Data File Preparation and Submission

The following guidelines apply specifically to researcher-sourced supplemental data, which researchers are responsible for formatting the data file.

Data Formatting

- **Acceptable formats:** SAS datasets (Version 9.4), Excel (xlsx), or CSV.
- **Variable names:** No spaces (e.g., FirstName not First Name).
- **File Naming:** Proposal Number + filename (e.g., UTA123_filename.xlsx).

Required Matching Data Elements

Include as much PII as possible to ensure successful matching. This PII will be used only for matching and will be removed before the processed data is transferred to the ERC.

- **Names:** First, middle, and last names must be in separate fields (letters only, no special characters)
- **SSN:** Must be 9 characters with no hyphens and include any leading zeros.
- **Date of Birth:** Must be in YYYYMMDD numeric format.
- **Additional TEA identifier (if available):** Unique ID, Student ID, Staff ID, TEAID (staff only).
- **Helpful non-PII:** Local Education Agency (LEA) identifier (6-digit district number) and campus identifier (9-digit campus number).

Data Transfer

After your dataset is prepared, contact your ERC director. Then, TEA/THECB will coordinate with a single designated representative from your research team to facilitate a secure data transfer.

- **Secure Transfer Portal:** All data files must be submitted through a secure transfer portal provided by TEA or THECB. **Do not send data via email.**
- **Consistency:** If submitting multiple files, they must all be in the same format.

- **Responsibility:** Researchers are responsible for formatting and preparing their data. TEA/THECB will **not** clean, reformat, or modify data to increase the matching rate, nor will they work with researcher-provided code.
- **File Rejection:** Data files that do not match the approved Data Dictionary (e.g., variables or values not in the approved data dictionary) will be rejected. The ERC director will be notified, and any amendments will require approval at the next Advisory Board meeting, placing the submission at the end of the processing queue.

Data Matching and Processing

After receiving and reviewing the submitted data file, the matching process will begin.

- **TEA Matching:** Matches individual-level data to the K-12 database, assigns a unique ID1, and forwards the data to THECB.
- **THECB Matching:** Matches the data to the higher education database, assigns a unique ID2, and removes all PII before transferring the de-identified, processed data to the ERC.

Processing Time

These are only estimates. Processing time will vary based on data quality, size, and scope. Submissions are processed in the order they are received.

- **TEA (ID1 Matching):** Expect 3-6 months.
- **THECB (ID2 Matching):** Expect an additional 2-3 months.

Important Notes

- Fuzzy matching will **not** be performed. Matches will not account for alternate spellings, nicknames, or abbreviations. Only exact matches will be performed based on SSNs or a precise combination of names and birthdates.
- TEA/THECB will **not** create variables, flags, or datasets, nor will they perform any data merging or linking for researchers.
- Data obtained through a Public Information Request **cannot** be processed as supplemental data. This data is masked and de-identified before release and cannot be reprocessed to add an ERC ID.

Updates and Additional Data

- **Adding New Variables:** To add new variables, you must submit an "Additional Data Request" and an updated Data Dictionary (including both pre-approved and newly requested variables). This request will be reviewed at a future Board Meeting. If approved, the supplemental data will be placed at the end of the processing queue. This rule applies even if the changes are minor.

- **Submitting Additional Years:** If your project was approved to include additional years of supplemental data, the processing queue for these years will begin at the end of the line.

Data Destruction and Access

- **Data Destruction:** TEA/THECB will destroy all supplemental data within **60 days** of the Advisory Board-approved project end date.
- **Access Restrictions:** Researchers are **prohibited** from accessing processed, de-identified supplemental data externally. All data must be accessed and used only within the secure ERC environment to protect its integrity and confidentiality.

Special Cases

Researchers with prior Board Approval

If your supplemental data was approved without a data dictionary or with a non-THECB template, you may be required to work with your ERC director to create and submit a dictionary using the current THECB template for review.

TEA/THECB/Publicly Available Datasets

If your project is approved to use TEA/THECB data as supplemental data, you must collaborate with the respective agency to submit the required Data Dictionary and Approval Form. For publicly available datasets, TEA/THECB may request a detailed Data Dictionary and additional information related to the allowed usage of the dataset.

Data Sharing Agreements

If you plan to merge supplemental data with TEA/THECB/TWC data within the ERC, you must secure a data-sharing agreement with all relevant state agencies (TEA, THECB, TWC) and your ERC. Coordinate with your ERC director to begin this process.

For all questions, researchers must contact their ERC directors.

Appendix A: Prohibited Data

The table below outlines categories and data elements that are prohibited from inclusion in supplemental datasets submitted to TEA and/or THECB. Data elements provided for matching purposes will be used solely for that purpose and will be removed before the supplemental data is transferred to the ERCs.

Prohibited Category	Examples of Prohibited Data Elements
Names & Personal Identifiers	Parents, Guardians, Teachers, Staff, Legal Names, Aliases, Mother's Maiden Name, Social Media Usernames, Physical Description (height, weight, distinct marks)
Direct Identifiers & Serial Numbers	SSN (outside matching keys), Driver's License Number, Client Number, Personal ID, Student ID, Financial Account Numbers, License Plate, IP Address, Vehicle ID/Serial Number, Certificate/License Numbers, Passport Number, Record IDs, Employer ID, Badge ID, Alternate IDs, Medical Record Numbers, Health Plan Numbers, Device Identifiers, Court Case Numbers, Researcher-created IDs
Relational & Family Identifiers	Marriage ID, Marriage License / Certificate Numbers, Sibling ID, Twin / Multiple Birth Identifiers, Family ID, Household ID
Individual & Residential Geographic Location	No variations of residential addresses are allowed. Home/Mailing Address, Place of Birth/Marriage/Death/Divorce, Geolocation Data (GPS coordinates), Postal Code (ZIP), Voting/Polling Location, Census Tract/Block/Block Group, Emergency Contact Address, Work Address. <i>(Note: County of Residence is permitted solely within THECB postsecondary data and strictly forbidden in TEA K-12 data).</i>
Dates & Timestamps	Specific Time-Based Identifiers: Birthdate (outside matching keys), Admission/Discharge Dates, Date of Death, Service Dates, Marriage/Divorce Dates, Conviction/Arrest/Disposition Dates, Event Participation Dates, Voting Dates, Employment Start/End Dates, Specific Timestamps
Contact Details	Phone Numbers, Fax Numbers, Email Addresses (personal/work), Job Title, Parent/Guardian Contact Information
Assessment & Biometrics	Item-level assessment responses, Raw Test Scores, Online Test Results, Voiceprints, Fingerprints, Retina/Iris Scans, Hand Geometry, DNA, Facial Recognition Data, Facial Photographs/Images

Health & Behavioral Health Data	Medical/Health Conditions, Genetic Data, Substance Use/Treatment Status, Disability Types, Clinical Notes, Mental Health Records, Prescription Records
Financial & Economic Data	Exact Income/Salary Amounts, Financial Data, Insurance Policy Numbers, Benefit/Compensation Information, Tax Return Line Items, Credit Scores, Property Asset Values
Employment & Workforce Details	Job Title, Direct Employer Names, Specific Work Addresses, Employer Tax IDs
Restricted Attributes	Demographic variables coded differently from existing ERC standardized definitions and allowable value categories; Religious Affiliation; Driver Records; Traffic Violations; Political Affiliation; Voting Party
Criminal Justice Records	Unique person identifiers, arrest IDs, offense IDs, State Identification (SID) numbers, or court case numbers Offense codes/degrees, charge levels, specific offense dates, incident timestamps, or domestic violence flags Date of arrest, arresting agency, booking numbers, or detention dates Court dispositions, appeal decisions, deferred adjudication status, or conviction dates Sentence length, probation/parole durations, suspended sentence lengths, court fines/costs

Additional Prohibitions:

- If a data element already exists in ERC data, researchers must use the ERC variable. Supplemental data that duplicates or overlaps with existing ERC data elements must use ERC definitions and allowable values.
- Open-ended survey responses or free-text fields may not be included in the data file.
- Derived or proxy variables intended to approximate any prohibited data element listed in the table above are strictly prohibited, regardless of how they are labeled.

Appendix B: Pre-Approved Aggregate-Level Data Sources

Default Rule

Pre-approval applies ONLY to aggregated, summary-level statistics or standard reference lookup tables. Public-use microdata, person-level extracts, household-level files, or individual-level records are NOT pre-approved, even when downloaded directly from a pre-approved agency website.

Pre-Approved Sources

For the sources below, researchers may use the level of aggregation the source agency has already published on its website or in its official reports. This does not include individual-level data, or any location information tied to individual residence.

Source	Approved Aggregation Level / Use
TEA AEIS (Academic Excellence Indicator System)/TAPR (Texas Academic Performance Reports)	Campus or district level, or summary level used in the agency's published report
THECB Accountability Data/Almanac	Institution level or summary level used in the agency's published report.
CCD (Common Core of Data)	School, district, state level, or summary level used in the agency's published report. Any location data is limited to where the building is, not where a person lives.
IPEDS (Integrated Postsecondary Education Data System)	Institution level, or summary level used in the agency's published report.
U.S. Census Bureau (ACS, Decennial)	State, regional, Metropolitan Statistical Area (MSA), county, census tract, block group, or block levels. Must be tied solely to school, campus, or institutional facility locations. Any use of census geographic units (including tracts, blocks, or block groups) tied to student, individual, or residential addresses is strictly prohibited. (<i>County of Residence is allowed exclusively within THECB postsecondary datasets</i>).
QCEW (Quarterly Census of Employment and Wages)	Statewide, MSA, or Large County totals by industry.

GIS/spatial data, TIGER, TEA campus/district spatial data	Boundary files (ISD, Region, Tract, Block Group) and point coordinates/street addresses are permitted solely for locating public infrastructure, schools, campuses, or institutional facilities.
Reference Lookup Table	Texas Common Course Numbering System (TCCNS), CIP Codes, NAICS/SIC, TCC, University Course Inventories, O*NET-SOC Occupation Classification
Inflation Rate / CPI (BLS)	Statewide or national economic index numbers.
College Scorecard	Institution or field-of-study/program level as published in official tools.

Note: While County of residence is accessible within THECB higher education administrative data, researchers are prohibited from using any residential or geographic location indicators to K-12 data. For projects utilizing both TEA and THECB datasets, county level residence should be evaluated solely within the postsecondary environment and cannot be linked across or merged into K-12 records.

Not Pre-Approved (Requires Advisory Board Review and Submission of Data Dictionary)

Examples (not exhaustive):

- Any individual-level or person-level data, regardless of source agency or public availability.
- Datasets requested at a more granular level than the pre-approved thresholds listed above.
- Institution-level or course-level Syllabus Data (unstructured text data can contain faculty/personnel details).
- Any dataset or variable not explicitly listed in the table above.

Mandatory Conditions for “Pre-Approved” Sources

- Use only the published version and aggregation level the source agency has already published on its site or in its official reports.
- All original small-cell suppressions, masks, or top-codes applied by the publishing agency must remain completely intact.
- If a data element already exists in ERC data, researchers must use the ERC variable.
- Researchers must list all pre-approved datasets and the aggregation level in their research proposal. Listing the pre-approved dataset on the approved proposal serves as the documentation requirement in lieu of submitting a separate Data Dictionary.

Appendix C: Examples of Personally Identifiable Information (PII) When Combined

The table below provides examples of how combining secondary identifiers with other data points can increase the risk of identifying individuals, highlighting specific combinations and their potential privacy implications.

Category	Secondary Identifiers	General Example of Potential PII	Example
Personal Characteristics	Race / Ethnicity	Geographic area + race/ethnicity + small district	Rare student demographic category in a small school district.
Language & Background	Primary Language	Geographic area + non-English language + school	Student speaking an infrequent non-English language at a specific local elementary school.
Educational Attainment	Degree / Credential Level	Geographic area + specialized advanced degree	Holder of a specialized doctoral/professional degree residing in a sparsely populated county.
Financial Information	Salary / Income Level	County + extreme salary tier	Individual with an extreme high-income outlier status within a rural district.
Affiliations	Religious / Political Affiliation	Specific affiliation + geographic unit	Rare religious or political affiliation indicated within a small municipality.
Health Attributes	Health Conditions	Specific health attribute + geographic area	Rare health condition documented within a small municipality.
Employment Details	Job Attributes	Geographic unit + specialized occupation	Infrequent professional occupation represented within a rural town.
Family / Household Composition	Household Size / Twin or Multiple Birth Indicator / Single Parent Status	Small school + rare household structure	A student designated as a triplet or belonging to a household with 8+ siblings within a single rural campus.