

Full length article

Knowing millions of students too well: High-entropy scores as deterministic quasi-identifiers for re-identification and data leakage in the Brazilian high school exam

Henrique Lindemann^a, Eder J. Scheid^a, Lisandro Z. Granville^a, Muriel F. Franco^{b,*}

^a Federal University of Rio Grande do Sul (UFRGS), Av. Bento Gonçalves 9500, Porto Alegre 91509-900, Brazil

^b Federal University of Health Sciences of Porto Alegre (UFCSPA), Sarmento Leite 245, Porto Alegre 90050-170, Brazil

ARTICLE INFO

Dataset link: <https://www.gov.br/inep/pt-br/acao-a-informacao/dados-abertos/microdados/enem>, <https://dadosabertos.mec.gov.br/sisu>, <https://github.com/ComputerNetworks-UFRGS/enem-microdata-privacy>

Keywords:

Data privacy
De-anonymization
Record linkage
Educational microdata
k-anonymity
Privacy–utility trade-off

ABSTRACT

Public microdata sit at the intersection of transparency mandates, data-protection law, and the technical realities of cybersecurity, while regulatory and judicial decisions often rely on threat models that underestimate modern re-identification capabilities. We study this tension in Brazil's National High School Exam (ENEM), where a court first mandated utility-preserving anonymization, then recognized substantial compliance through a separated architecture that forgoes core analytical utility. We show that precise candidate-level performance scores are high-entropy quasi-identifiers that adversaries can use to link identities across complementary public releases. Across 57.7 million records from fifteen editions (2009–2023), edition-level singleton rates average 99.9985%. On commodity hardware, an exact join of files downloadable from official portals uniquely matches 26.59 million of 26.61 million candidate-year records in the federal admission system (99.91%). Cross-year canonicalization estimates 14.69 million distinct individuals with linked records. These links associate civil identities with socioeconomic data such as family income and household assets, and a smaller channel conditionally exposes accessibility-related proxies for disability. Moreover, no tested score generalization removed record individualization without degrading ranking and group-gap estimates. Multidimensional performance measures should be treated as quasi-identifiers, not just as analytical attributes.

1. Introduction

The Brazilian National High School Exam (ENEM) is a large-scale assessment and a principal gateway to public higher education through the Unified Selection System (SISU) (da Silveira et al., 2015). Its microdata, released by the National Institute for Educational Studies and Research (INEP), support research and public-policy evaluation (Franco et al., 2024) by historically providing participant-level performance alongside demographic, educational, and socioeconomic attributes (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2024). This analytical value also creates a disclosure problem: removing direct identifiers does not prevent an ostensibly anonymous record from becoming attributable when released attributes can still be linked to an identified person. This risk is not specific to Brazil; it arises wherever granular, multidimensional performance data are released

alongside an obtainable source that ties the same values, or a derivable equivalent, to identities.

ENEM's microdata disclosure policy has consequently moved through three architectures: granular publication, retroactive suppression, and, for ENEM 2024, the separation of participant attributes from examination results without a shared linkage key. Each step traded analytical value for protection, and the 2024 separation makes that trade explicit: it prevents a candidate's socioeconomic profile from being joined to their performance within the released architecture, but for the same reason it interrupts analyses that require both in the same unit. Two successive judicial decisions placed this privacy–utility tension at the center of a dispute (Justiça Federal da 1ª Região, 2025, 2026).

A prior privacy assessment of Brazilian educational microdata established the disclosure power of demographic and geographic quasi-

* Corresponding author.

E-mail addresses: hlindemann@inf.ufrgs.br (H. Lindemann), ejscheid@inf.ufrgs.br (E.J. Scheid), granville@inf.ufrgs.br (L.Z. Granville), muriel.franco@ufcspa.edu.br (M.F. Franco).

<https://doi.org/10.1016/j.cose.2026.105080>

Received 9 June 2026; Received in revised form 21 July 2026; Accepted 26 July 2026

Available online 30 July 2026

0167-4048/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

identifiers (Alvim et al., 2022, 2020). Its findings motivated the suppression of some fields in ENEM microdata, including school, residence, and birthplace information (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022b). Candidate performance, however, remained primarily an analytical outcome. INEP's impact assessment retained scores and item-level responses within the processing scope, but its mitigation centered on geographic and demographic attributes without treating scores as linkage keys (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022a, Secs. 3.1.10–3.1.28 and 3.2.1). This leaves a clear blind spot. We argue that precise multidimensional scores are themselves highly distinctive quasi-identifiers and can support deterministic linkage when an obtainable auxiliary source associates the same values, or a derivable equivalent, with identity.

In this work, we evaluate this hypothesis on the ENEM microdata through three complementary analyses. (i) We characterize score distinctiveness under eight adversary models, ranging from score-only knowledge to combinations of scores with demographic and geographic context. We first test its persistence through a longitudinal k -anonymity audit across fifteen editions (2009–2023), and then examine 2023 in detail through entropy and individualization analyses. (ii) We test how individualization becomes identity association through three linkage vectors with different auxiliary sources and capabilities. (iii) Finally, we assess whether fixed score generalization, a statistical disclosure control (SDC) technique, can remove record singularity without degrading analytical utility.

The results show that score distinctiveness persists across cohorts. Under the exact five-score tuple model, an edition-level mean of 99.9985% of complete-score records belonged to singleton equivalence classes across the fifteen editions (Section 6.1). The three linkage vectors then demonstrate increasingly direct paths from score-based individualization to identity association. Most notably, the exact linkage between complementary federal releases uniquely matches 26,589,382 of 26,612,533 candidate-year records in the federal admission system (99.91%), while cross-year canonicalization estimates 14,688,801 distinct individuals affected (Section 6.3.3). This exact join, like the other evaluated attacks, ran on commodity hardware without access to a breached system.

Once identity linkage is established, socioeconomic attributes—including family income, housing conditions, and household assets—become directly associated with named individuals (Section 6.5). Exambook codes assigned to adapted booklets create a distinct conditional channel that can expose accessibility-related proxies for disabilities once such linkage succeeds (Section 6.5.1). The source releases underlying this exposure remained publicly downloadable at the study's July 15, 2026, cutoff, after the coordinated disclosure described in Section 8. Our SDC experiment, in turn, found that no tested score resolution eliminated singleton records under the most-informed adversary while keeping the four evaluated utility outcomes near their original values.

The main contributions of this work are as follows:

- We show that precise multidimensional performance scores are high-entropy quasi-identifiers in the 2023 ENEM release and that their near-uniqueness persists across fifteen editions (2009–2023).
- We evaluate the transition from statistical individualization to identity linkage through three linkage vectors with distinct auxiliary sources, capabilities, and limitations.
- We quantify the privacy–utility profile of fixed score generalization and, together with the identified reconstruction and linkage paths, show that granular performance data published alongside obtainable identity sources must be assessed within their full release environment.

The remainder of the paper is organized as follows. Sections 2–4 develop the policy context, privacy model, and score mechanism; Sections 5 and 6 present the study design and evidence; and Sections 7–9 address implications, ethical disclosure, and conclusions.

2. ENEM microdata architecture and policy conflict

ENEM microdata have been released under three architectures, referred to throughout this article as Phases 1–3. Phase 1 denotes the original relational releases for 2009–2019. Because the Phase 2 architecture was later applied retroactively to those historical packages, Phase 1 cannot be fully characterized from the files currently hosted by INEP. We therefore reconstruct its field structure from Willrich's documentary comparison of the official ENEM data dictionaries, which tracks variable introduction, renaming, recoding, and removal across the series (Willrich, 2022, Sec. 2.3), and corroborate that reconstruction with INEP's own account of the suppressed fields (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022b). In those original releases, precise scores appeared alongside detailed demographic, socioeconomic, geographic, school, and sensitive specialized-assistance information; fields varied by edition.

INEP presented Phases 2 and 3 as compliance measures under Brazil's General Data Protection Law (LGPD), but they changed different aspects of the release. Phase 2, referred to by INEP as the Simplified Model, was introduced for ENEM 2020 and subsequently applied retroactively to earlier packages. It retained the relational structure but reduced participant attributes by removing the school code and the municipalities of residence and birth, grouping age, and dropping specialized-assistance fields (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022a, Secs. 3.1.11–3.1.28), Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2022b). Phase 3, introduced for ENEM 2024 and not applied retroactively, instead changed the release structure. To restore the school code in response to user demand while preserving anonymity, INEP divided the release into two files (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2025, pp. 6–7). PARTICIPANTS contains participant attributes and the socioeconomic questionnaire, whereas RESULTS contains school and performance data. The files share no linkage key, which INEP states prevents school information from being used to re-identify participants. The same separation also prevents socioeconomic profiles and performance from being joined within the published release, reducing its relational utility. Table 1 summarizes the chronology and principal field-level changes.

These changes unfolded within a policy conflict between access to educational information and protection against disclosure. Law No. 15,017 of November 2024 made the transparency principles of the Access to Information Law (Brazil, 2011) explicit for educational data. It requires data and microdata from national educational exams (such as ENEM) to be publicly available, provided that their treatment, disclosure, and sharing occur in anonymized form whenever possible (Brazil, 2024, Art. 1). In a Public Civil Action that preceded that law, the Federal Public Prosecutor's Office (MPF) challenged INEP's Phase 2 suppressions, arguing that the re-identification scenario INEP assumed—an adversary manually retrieving birthplace, residence, age, and school information from a target's open social-media profile—fell outside what the MPF considered reasonable and did not justify withholding information needed to diagnose educational conditions and inform public policy (Ministério Público Federal, 2022, pp. 14–17). The October 8, 2025 merits judgment in that action (Decision A) ordered restoration of the suppressed microdata through utility-preserving anonymization within 180 days (Justiça Federal da 1ª Região, 2025, Sec. 5 and dispositive items (a)–(c)). A later judgment on the motions for clarification (Decision B), however, recognized substantial compliance through the separated Phase 3 architecture and revoked the deadline (Justiça Federal da 1ª Região, 2026, Secs. 1 and 2.3; dispositive). Structurally, the accepted architecture forgoes precisely the relational utility that the initial action and Decision A sought to preserve.

The privacy analysis preceding these decisions treated combinations of demographic attributes, geographic information, and school affiliation as quasi-identifiers. The audit commissioned by INEP showed that a few easily obtained attributes could deterministically re-identify a

Table 1
ENEM disclosure architectures and chronology. Fields vary by edition.

Phase/architecture	Chronology	Relational structure	Principal documented change
Phase 1: Original relational releases	Original ENEM 2009–2019 releases; reconstructed from documentary sources and not accessed in this study.	Performance and participant attributes appeared in the same candidate-level record.	The historical dictionaries document granular demographic, socioeconomic, school, accessibility-related, and geographic content (e.g., residence, birthplace, and examination location), with edition-specific schemas (Willrich, 2022, Sec. 2.3).
Phase 2: Simplified model	Introduced for ENEM 2020; used for subsequent editions through 2023 and applied retroactively to earlier packages.	Performance and participant attributes remained in the same candidate-level record.	School code, residence/birthplace, and specialized-assistance fields were removed, and age was grouped; state and municipality of examination remained available (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022b), Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2022a, Secs. 3.1.11–3.1.28).
Phase 3: Separated release	ENEM 2024 architecture; not applied retroactively.	PARTICIPANTS and RESULTS have no shared linkage key in the published files.	The school code removed in Phase 2 was restored and appears with results; socioeconomic attributes and performance can no longer be joined within the released architecture (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2025, pp. 6–7).

large share of students in the Basic Education Census (Alvim et al., 2020, Table 4.3.7, p. 43). Yet the audit assessed such triangulation, not academic performance, and ENEM mitigation retained the same focus. It suppressed school and residential identifiers and grouped age while retaining examination geography. Precise objective-test scores, essay-component scores, and item responses also remained within the processing scope but were not treated as linkage keys (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022a, Secs. 3.1.10–3.1.28 and 3.2.1). Neither the audit nor the mitigation evaluated multidimensional performance as a cross-release linkage key. This is the blind spot we examine.

3. Related work and privacy model

Removing names and other direct identifiers does not make a microdata release unlinkable. Ordinary attributes can act as quasi-identifiers when they are sufficiently distinctive and are available, alone or in combination, in another source. Sweeney’s early analysis showed that combinations of demographic attributes can distinguish large portions of a population even when none is a direct identifier (Sweeney, 2000); her subsequent k -anonymity model formalized a protection requirement against this linkage condition by requiring each released record to share its quasi-identifier values with at least $k - 1$ others (Sweeney, 2002). The relevant unit of analysis is not a dataset in isolation, but the set of complementary releases through which an adversary may obtain the same attributes or an equivalent linkage key.

Three outcomes must nevertheless be kept separate. *Individualization* means that a record is distinguishable within the released dataset under a specified projection; in the model used here, a singleton equivalence class has $k = 1$. *Identity linkage* additionally requires an auxiliary source that associates the observed key, or a derivable equivalent, with a named person. *Attribute disclosure* occurs when information about that person can then be learned from the linked record or even inferred from a non-singleton class. These are related but non-equivalent outcomes: a release may contain many singleton records without an accessible identity source, while a non-singleton class may still reveal a sensitive value shared by its members. Thus, $k = 1$ diagnoses a necessary statistical condition for the exact linkage attacks evaluated in this study, but it does not by itself establish that identity has been recovered. Conversely, $k > 1$ means that the tested key did not individualize the record; it is not a general guarantee against disclosure (Sweeney, 2002). Section 5 defines the equivalence classes and diagnostics calculated in our experiments.

De-anonymization experience also shows that linkage depends on an adversary’s auxiliary knowledge and its overlap with the target data. The 2006 AOL search-log release illustrated the point beyond conventional demographic data. Pseudonymous user numbers did not prevent journalists from re-identifying one user through the distinctiveness of her accumulated query history (Barbaro and Zeller, 2006). Narayanan and Shmatikov later showed that sparse, high-dimensional

rating data could be de-anonymized even when auxiliary knowledge was incomplete or imprecise (Narayanan and Shmatikov, 2008). Together, these examples place the availability and precision of external knowledge within the threat model, rather than treating risk as a property of the released dataset alone. In the closest Brazilian educational antecedent, Alvim et al. modeled disclosure attacks on microdata from INEP’s Educational Censuses using combinations of demographic attributes, geographic information, and school affiliation (Alvim et al., 2022, 2020). Their experiments established the power of such combinations at scale, but did not evaluate precise multidimensional ENEM performance as a cross-release linkage key.

Protecting a release, however, carries its own cost: the transformations that reduce linkage risk also alter the data on which analyses depend. In the HarvardX and MITx release (Daries et al., 2014), k -anonymization disproportionately suppressed records from smaller courses and more active participants, changing activity and certification measures as well as correlations used in downstream analyses. Utility must therefore be evaluated against the specific analyses a release is intended to support. For ENEM, performance variables are central to analyses of educational inequality, in which scores are examined alongside family income, parental education, race, school sector, and geography (Franco et al., 2024; da Silveira et al., 2015). ENEM reports separate proficiencies based on Item Response Theory (IRT) for its four objective areas (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2021). The production of these scores is documented, but two disclosure questions remain. How often do the published values and their multidimensional tuples repeat within an edition and across the disclosure series? Can an auxiliary source associate those values, or a derivable equivalent, with identity? Section 4 develops the score-specific mechanism; the subsequent audit and linkage experiments test its persistence and operational consequences.

4. High-entropy scores: Distinctiveness and linkage mechanism

ENEM releases five performance values for each record. Four are objective-area proficiencies estimated through a three-parameter IRT model, which combines each candidate’s response pattern with calibrated item discrimination, difficulty, and guessing parameters to produce values on a continuous scale (Baker, 2001; Andrade et al., 2000; Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2021). The fifth value is the separately produced essay score and is not part of the IRT output. We characterize two score representations as candidate linkage keys: a derived arithmetic mean of the five scores, rounded to two decimals, and the exact five-score tuple at its original microdata precision. Their empirical distributions determine how strongly they distinguish records; as quantified below, both carry substantially more entropy than the demographic and geographic attributes examined.

Table 2

Empirical entropy of ENEM 2023 attributes and score keys. The analysis population contains $N = 2,678,264$ records with nonmissing values for all five scores. For each row, M and H are computed over the observed nonmissing values of that attribute or key.

Attribute or key	Distinct values (M)	H (bits)
Gender	2	0.96
Race/Color	6	1.63
Age range	20	3.20
State of examination	27	4.27
Municipality of examination	1,715	9.23
Derived arithmetic mean	31,849	14.25
Exact 5-score tuple	2,678,235	21.35

The equations below show how to represent, for a discrete variable X , the self-information of an observed value x and the Shannon entropy (Shannon, 1948) of the distribution.

$$I(x) = -\log_2 p(x) \quad H(X) = -\sum_x p(x) \log_2 p(x)$$

Thus, entropy is the expected self-information, $H(X) = \mathbb{E}[I(X)]$. Self-information distinguishes common from rare realizations, whereas entropy summarizes their distribution. In the 2023 complete-score population ($N = 2,678,264$), for example, 409,480 records have state of examination SP and 6070 have RR. Their self-information values are 2.71 and 8.79 bits, respectively. The 6.08-bit difference corresponds to a 67.5-fold contrast in the resulting record pools, illustrating how geographic sparsity increases distinctiveness.

Table 2 reports the empirical distributions of the seven attributes and keys used in the 2023 analysis.

The contrast supports the characterization of score representations as high-entropy quasi-identifiers. The derived mean carries 14.25 bits across 31,849 observed values, while the exact tuple carries 21.35 bits across 2,678,235 values; the categorical and geographic attributes range from 0.96 to 9.23 bits. The tuple is consequently almost injective: among the $N = 2,678,264$ records, 2,678,225 (99.9985%) belong to singleton classes and only 39 share their tuple. Its entropy approaches $\log_2 N$ because nearly all records occupy singleton tuples and M is almost N . A near-injective shared key enables exact joins across complementary releases, making this distinctiveness the structural precondition for population-scale linkage: when an obtainable source associates the same tuple, or a derivable equivalent, with identity, individualization becomes identity linkage.

To measure how much auxiliary knowledge individualizes a record, we define eight adversary models spanning a range of contextual knowledge (Table 3). The first six form a graded scale over the derived mean, from an adversary holding only a published average score (Stranger) to one who also knows the target's gender, municipality, and age range (Neighbor). The names reflect this gradient: a Stranger sees only the public score, whereas a Neighbor also knows where a candidate took the exam and their age. The Specialist and Profiler instead assume access to the exact five-score tuple, the latter adding the same demographic and geographic context. Throughout, state and municipality refer to the place of examination, the only geography retained in the Phase 2 microdata.

The scenarios deliberately omit other edition-specific fields present in the analyzed releases, including socioeconomic questionnaire responses. For a fixed record population and missing-value convention, appending an attribute to a linkage key can only preserve or split its equivalence classes, never merge them. The scenario-specific singleton rates reported in Section 6.2 are therefore conservative: an adversary who also knew those omitted fields could only individualize at least as many records, never fewer.

Table 3

Adversary models grouped by linkage-key family. Each row gives the complete key used to form equivalence classes.

Adversary	Assumed knowledge/complete key
Derived-average family	
Stranger	Arithmetic mean
Acquaintance	Mean + gender
Statewide	Mean + gender + state of examination
Regional	Mean + gender + state of examination + age range
Local	Mean + gender + municipality of examination
Neighbor	Mean + gender + municipality of examination + age range
Exact-tuple family	
Specialist	Exact five-score tuple
Profiler	Tuple + gender + municipality of examination + age range

5. Methodology: Score linkage

Building on the high-entropy property established in Section 4, we test whether score distinctiveness persists across editions and materializes as linkage under real auxiliary-data conditions. The study comprises a longitudinal k -anonymity audit of the 2009–2023 ENEM series, a detailed 2023 snapshot, three empirical linkage vectors, and an evaluation of score generalization against privacy and analytical-utility outcomes.

Our primary diagnostic is k -anonymity (Sweeney, 2002). For each adversary scenario, records sharing the same linkage key form an equivalence class of size k . A record is individually distinguishable when it forms a class of one ($k = 1$).

A supplementary verification package accompanies the manuscript. It provides the sanitized aggregate outputs, schemas, pseudocode, and synthetic examples needed to audit the reported claims without access to the operational code or candidate-level data. The package is organized by evidence family; specific families are cited where relevant in the text. Access information and a description of its contents are provided in the Data Availability statement.

5.1. A Longitudinal k -Anonymity Audit (2009–2023)

We audit the 15 ENEM editions from 2009 through 2023 as currently hosted under INEP's Phase 2 disclosure model, including releases retroactively subjected to the geographic and demographic suppressions described in Section 2. From 86,110,930 enrollment records, we retain the 57,708,475 candidate records with nonmissing values for all five scores. The restriction serves comparability, not linkage feasibility: it holds the score representation fixed so that one set of adversary models applies across all fifteen editions. The reported uniqueness rates are therefore conditional on the complete-score population.

For each edition, we instantiate the scenarios of Table 3. The average-based adversaries use the arithmetic mean of the five scores as an institution-agnostic proxy for the program-specific weighted means used in real admissions, rounding it to two decimals; the tuple-based adversaries instead use the exact five-score tuple at INEP's published one-decimal precision. For each year and scenario we report the share of complete-score records in singleton classes ($k = 1$), isolating whether the entropy-predicted distinctiveness is a cohort anomaly or a persistent property of the series.

Annual changes in schema, column casing, and variable names for the five scores, gender, age range, municipality of examination, and state of examination are harmonized. Missing or unreported demographic values are preserved as explicit categories and form their own equivalence classes. The corresponding field definitions are documented in the supplement's schema catalog.

Table 4

Empirical linkage vectors. The success criterion is evaluated against ENEM microdata; limitations qualify the scope or certainty of each vector.

Vector	Auxiliary source	Available knowledge/adversary model	Linkage key	Regime	Success criterion	Principal limitation
A	UFRGS admission portal and published program-weight rules	Name, program, quota category, weighted mean; Stranger–Neighbor knowledge added by scenario	Recomputed program-specific weighted mean plus filters	Deterministic pivot; conditional when inferred filters are used	One ENEM record remains under the stated scenario ($k = 1$)	Single-institution case study; inferred demographic and quota attributes are subject to misclassification
B	Public self-disclosure or third-party score platforms	Identity context and five-score tuple; Specialist analogue	Exact five-score tuple	Exact deterministic	The disclosed tuple retrieves one ENEM record ($k = 1$)	Conditional on disclosure or access; platform records were not accessed and population prevalence is not estimated
C	MEC (Brazilian Ministry of Education) SISU Open Data, 2010–2023	Civil identity, partial Brazilian individual taxpayer identification number (CPF), birth data, and five-score tuple; Specialist analogue	Exact five-score tuple	Exact deterministic	A SISU candidate-year retrieves one ENEM record ($k = 1$)	Limited to SISU applicants; cross-year person counts depend on deduplication

5.2. ENEM 2023 snapshot

We use the 2023 cohort, the final edition released under the Phase 2 architecture, to examine how the longitudinal vulnerability arises within a single contemporary cohort ($N = 2,678,264$ complete-score records). We add the quasi-identifiers of Table 3 sequentially, recomputing $k = 1$ at each stage to isolate the marginal identifying contribution of each attribute added to the score-based key.

The population average may conceal uneven risk. High-performing candidates occupy the sparse upper tail of the score distribution, where fewer records share similar performance, so we repeat the audit for the full cohort and for the top 1% and top 0.1% ranked by mean score. This isolates whether proximity to the score ceiling raises distinctiveness before any demographic attribute is added.

Candidate-pool size provides a second source of heterogeneity. Knowledge of the municipality of examination partitions the national cohort into groups containing different numbers of complete-score candidates, so the same score information may be more identifying where fewer candidates took the exam. We compute Local and Neighbor uniqueness separately for each municipality and relate these rates to the number of complete-score candidates observed there. The Local scenario combines average score, gender, and municipality, while Neighbor additionally includes age range; their difference measures the marginal contribution of age within each municipal candidate pool.

5.3. Empirical validation: Linkage attack vectors

As documented by Alvim et al. (2022) in their technical interactions with INEP, purely theoretical evidence of vulnerability is sometimes underestimated by public managers, who rely on a “safety in numbers” fallacy and dismiss findings from other datasets as inapplicable to their own. The authors report that perception shifts only when concrete attacks are reproduced on the agency’s own releases, which in turn motivates policy change. Following this lesson, this stage examines whether that distinctiveness materializes as linkage under real auxiliary-data conditions.

We evaluate three linkage vectors, each drawing on a different external source that pairs ENEM scores with identity. *Institutional linkage* (Vector A) targets university admission lists, which publish candidates’ names alongside program-weighted scores. *Self-disclosure linkage* (Vector B) targets scores that candidates themselves expose—on social media or through third-party simulators—as a complete five-score tuple. *Governmental linkage* (Vector C) targets the federal SISU Open Data,

which already pairs civil identities with exact score tuples in a single release.

The three form a gradient of operational friction, the effort required to convert score distinctiveness into an identity match. Institutional linkage demands the most: assembling institution-specific admission data, recovering program weights, reconciling heterogeneous formats, and recomputing the published weighted score before matching. Self-disclosure linkage requires only locating a public disclosure, after which the five-score tuple matches directly. Governmental linkage demands the least, because the federal files already consolidate civil identities and exact score tuples, reducing linkage to a direct database join. Vector A therefore serves as the high-effort endpoint, since a single institution suffices to show that linkage survives even a costly reconstruction, while Vector C serves as the low-effort endpoint, where a direct join requires no reconstruction. The regimes differ accordingly: some institutional scenarios rely on inferred or assumed quasi-identifiers and are conditional on them, whereas the other two match the disclosed tuple at its original precision. Table 4 summarizes each vector’s auxiliary source, adversary knowledge, linkage key, regime, success criterion, and limitation.

5.3.1. Vector A: Institutional linkage

Vector A represents the most complex scenario as it requires reconstructing the score linkage key. Institutions in SISU publish their own selection results through heterogeneous interfaces, including PDF lists, institutional portals, and interactive systems. The 2024 SISU cycle included 127 institutions (*Secretaria de Comunicação Social, 2024*), and the variables disclosed and their presentation are not uniform. Vector A examines the conditions under which an institutional disclosure containing a nominal result and a program-specific weighted mean can be linked back to ENEM microdata when the corresponding weighting rules are also public.

We operationalize this vector as a case study combining the 2024 admission cycle at the Federal University of Rio Grande do Sul (UFRGS)—9455 candidates across 100 programs—with the 2023 ENEM microdata. A preliminary manual inspection of admission portals from other major universities (e.g., UFMG, UFRJ, and UFCSPA) identified comparable disclosures associating nominal results with exact weighted scores. This inspection motivated the broader threat model, but neither it nor the UFRGS case study constitutes a systematic survey or estimates prevalence across SISU institutions.

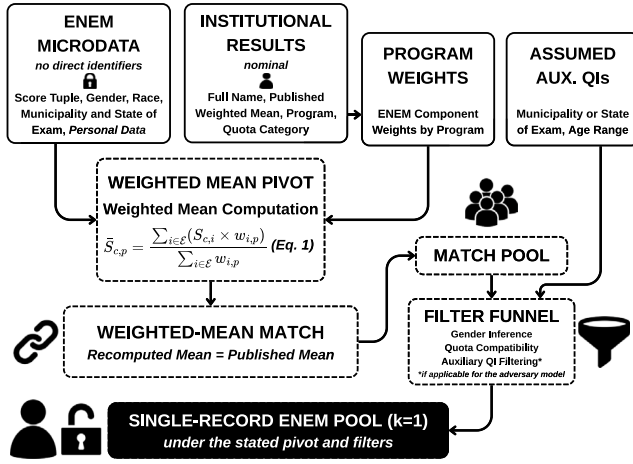


Fig. 1. General Vector A pipeline, instantiated with UFRGS data. The reconstructed program-specific mean forms an initial ENEM match pool, which inferred and auxiliary quasi-identifiers may further narrow.

Data sources and weighted mean pivot. The concrete UFRGS implementation integrates four inputs:

1. **ENEM microdata:** the target records, containing the five scores, the demographic and geographic variables used in the adversary scenarios, and the school and socioeconomic attributes associated with a record after linkage.
2. **Nominal institutional results:** the external source associating candidates' identities with their programs and published weighted means. In the UFRGS case, the admissions portal provides each candidate's name, program, quota category, and weighted mean.¹ The university's PDF admission lists omit the weighted mean and are therefore not the score source used in this analysis.
3. **Program weighting rules:** the weights needed to reconstruct the institutional score from the five ENEM components. Ordinance No. 21/2012 of the Ministry of Education, Art. 5(IV), requires participating institutions to define and communicate these weights (Ministério da Educação, 2012). At UFRGS, they are established by a public resolution of the Teaching, Research, and Extension Council (CEPE).² We manually map the published weights to each program in the admission list, validating ambiguous name matches by hand.
4. **Auxiliary quasi-identifiers (QIs):** gender, state or municipality of examination, and age range, introduced according to the adversary scenario. These attributes are not disclosed by the UFRGS portal and are treated as inferred or assumed knowledge.

Fig. 1 summarizes the pipeline. For every ENEM record and program p , we reconstruct the weighted mean

$$\bar{S}_{c,p} = \frac{\sum_{i \in \mathcal{E}} (S_{c,i} \times w_{i,p})}{\sum_{i \in \mathcal{E}} w_{i,p}} \quad (1)$$

where $\mathcal{E}$ is the set of five ENEM exams, $S_{c,i}$ is candidate c 's score for exam i , and $w_{i,p}$ is the weight assigned to that exam by program p . Thus, $\bar{S}_{c,p}$ is candidate c 's reconstructed weighted mean for that program. The recomputed means are formatted to the two decimals published

by the portal and indexed by program. Each nominal candidate is assigned the equivalence class of ENEM records sharing that program and weighted mean; $k = 1$ denotes a unique candidate pool under the pivot. We test round-to-even, half-up rounding, and truncation, as well as decimal-precision and mean-tie sensitivity.

Inference and auxiliary filters. The pivot is evaluated alone and with filters that model increasing adversary knowledge. Gender is inferred from first names using *genderBR*,³ which derives statistical name-gender associations from the 2010 IBGE Census. Quota categories published by the portal are used as compatibility heuristics against self-declared ENEM attributes. Public-school categories exclude records marked as private school. Race-targeted categories for Black, Brown, and Indigenous candidates, known in SISU as PPI, exclude records self-declared as White or Yellow/Asian. Public-school categories without the PPI designator apply an inverse non-PPI heuristic and exclude Black, Brown, and Indigenous declarations. To separate linkage produced by the weighted-score pivot from additional narrowing introduced by inference, we evaluate the pipeline with and without the gender and quota filters in a full ablation. These filters are treated as heuristics, not as ground-truth demographic classifications.

Filter uncertainty and fallback. The compatibility filters depend on information recorded independently in the SISU and ENEM systems. For race/color specifically, INEP's impact assessment confirms that the ENEM field is self-declared (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022a, Sec. 3.1.21). The SISU quota category and the separately recorded ENEM declaration may therefore be inconsistent. Such differences may exclude the true ENEM record or leave another record as an apparent singleton. Accordingly, $k = 1$ after inferred filtering is not, by itself, a validated nominal association. When strict quota filtering produces an empty pool, the pipeline progressively removes the inverse non-PPI heuristic and then the PPI compatibility check, while retaining the public-school filter. These relaxations enlarge the pool and provide a more conservative uniqueness estimate. Their frequency and effect are reported in the ablation.

Auxiliary-QI uncertainty bounds. Because the auxiliary QIs are scenario assumptions, we evaluate the possible values specified by each scenario. Statewide is calculated across observed states, Local across observed municipalities, Regional across state and age-range combinations, and Neighbor across municipality and age-range combinations. For each scenario, we report k_{min} , k_{median} , and k_{max} , corresponding to the worst, typical, and best privacy outcomes across the evaluated possibilities.

Limitations of vector A. Vector A has four principal limitations. (a) *Representativeness:* the empirical analysis covers one institution and admission cycle. The exploratory inspection of UFMG, UFRJ, and UFCSPA motivated the general threat model but was not a systematic survey of SISU institutions. (b) *Gender inference:* name-based classification can misclassify uncommon, ambiguous, or foreign names and is sensitive to regional naming patterns. Moreover, the ENEM field used for matching is binary and does not represent non-binary identities. (c) *Quota compatibility:* race/color declarations are collected separately in SISU and ENEM and may differ, while missing or non-standard school categories limit the public-school filter. The inverse non-PPI rule is a heuristic and does not establish a candidate's racial identity. (d) *Auxiliary QIs:* state, municipality, and age-range results are scenario bounds rather than observations of each candidate's attributes.

¹ As of July 15, 2026: <https://www1.ufrgs.br/PortalEnsino/GraduacaoProcSeletivo/index.php/DivulgacaoDadosChamamento>.

² As of July 15, 2026: <https://www.ufrgs.br/cepe/wp-content/uploads/2022/08/Res-56-Tabela-Pesos-provas-ENEM-SISU-5-Alterada-pelas-Res.-25-26-27-e-28-2021-5a-Versao-WORDPRESS-1.pdf>.

³ <https://github.com/meirelesff/genderBR>.

5.3.2. Vector B: Self-disclosure linkage

Vector B examines what occurs when a candidate's complete five-score tuple circulates outside the ENEM microdata. Unlike Vector A, the adversary does not need to recover program weights, reconstruct a weighted mean, or infer demographic attributes. Once the tuple is associated with identity context, it can be searched directly in the microdata. Vector B therefore tests individual-level linkage conditional on an identity-associated tuple.

Scope and procedure. We conducted a qualitative, documentary review of two disclosure channels. First, we manually searched X, Instagram, TikTok, and Reddit in late 2025 for public posts and screenshots containing ENEM scores. Second, we inspected the publicly visible input requirements of third-party SISU simulation platforms to determine whether candidates are asked to provide scores together with account or contact information. We did not automate collection from social media, access platform databases, or retain a corpus of disclosed tuples. This procedure identifies possible disclosure routes but does not estimate their prevalence.

Exact tuple lookup. Given an identity-associated five-score tuple, the adversary matches it exactly against the corresponding ENEM edition. The participant portal and the public microdata report the five scores at one-decimal precision. A tuple retrieving one record ($k = 1$) satisfies the success criterion for an individual association; a collision ($k > 1$) remains ambiguous, and no match ($k = 0$) is unsuccessful.

Limitations of vector B. Vector B depends on an external disclosure and does not support a national exposure estimate. Public posts are fragmented, may be deleted, and may lack sufficient identity context. Platform-mediated exposure remains conditional because their internal records were not accessed.

5.3.3. Vector C: Governmental linkage

Vector C tests exact linkage between two complementary federal releases. The SISU Open Data files, published by Brazil's Ministry of Education (MEC), associate civil identity fields with the five ENEM performance scores, whereas INEP's historical microdata contain the same score tuple without names. We link 14 annual SISU cohorts from 2010–2023, assembled from 28 semester-level selection cycles, to the corresponding ENEM editions from 2009–2022. For each cycle, the linkage corpus uses the regular-admission files, which publish the complete candidate ranking for every program.

Preparation and candidate-year linkage. Because the SISU release format changed over time, we harmonize annual schemas, delimiters, encodings, and column names. Names are case-folded, stripped of accents, punctuation-normalized, and whitespace-collapsed; rows whose raw name contains digits are excluded as corrupted. Because some SISU editions publish only the birth year, we reduce all birth dates to the year for cross-cohort consistency. Records missing or reporting zero in any of the five scores are excluded on both sides. SISU scores are parsed from Brazilian comma-decimal strings or already-numeric fields, whereas ENEM scores are read numerically; values on both sides are then standardized to the published one-decimal precision and stored in the same numeric representation before lookup. No tolerance-based or approximate score matching is used.

SISU may contain several rows for one applicant, including program choices. We collapse these rows within each annual cohort using the composite key

$$Key_{\text{dedup}} = \text{normalize}(\text{Name}) \mid \text{BirthYear} \mid \text{CPF}_{4-8}$$

The five CPF digits are chosen for consistency. Among the mask-affected files documented in the supplement, only one partition falls within the *Chamada Regular* corpus used here—the first semester of 2017—which exposes positions 3–8 rather than the standard 4–9. We use positions 4–8, which are common to both formats. Each surviving

key defines one candidate-year, which is linked by looking up its exact five-score tuple in the corresponding ENEM edition. Civil attributes do not enter this join. One retrieved record ($k = 1$) is a successful exact linkage, multiple records ($k > 1$) are ambiguous, and a lookup retrieving no record is recorded as $k = 0$.

Longitudinal person estimate. The linkage outcome is measured at the candidate-year level, but estimating how many people are exposed requires collapsing applicants who recur across editions. Exact name normalization alone can split the same person when spelling changes. We therefore compare names only within groups sharing birth year and CPF fragment. Initials, generational suffixes, name particles, middle-name omissions, and Levenshtein distances of at most six with first-name consistency are accepted as spelling variants; other pairs are rejected. The union-find algorithm consolidates connected variants into one canonical name before the composite key is hashed for persistent cross-year tracking.

To quantify the sensitivity of the person estimate to this modeling choice, we rerun the pipeline without canonicalization and with three coarser keys: name plus CPF fragment, name plus birth year, and name alone.

Limitations of Vector C. The principal limitation concerns longitudinal person counting. Without civil-registry identifiers, legal name changes and unmodeled variants can split one person across keys and inflate the estimate, whereas homonyms or erroneous similarity merges can collapse different people and reduce it. Because these errors act in opposite directions, we report the full-key result as an estimate with sensitivity analyses.

5.4. Statistical Disclosure Control (SDC) defenses

The granularity of the five-score representation is the main empirical driver of the observed identifiability. We therefore evaluate score generalization through interval recoding. Generalization is a standard non-perturbative SDC strategy that reduces quasi-identifier resolution while retaining a deterministic and interpretable representation of the underlying values (De Capitani di Vimercati et al., 2023), a relevant property for public educational assessments. By coarsening the published performance representation, we test the resulting tension between privacy and utility: whether mapping scores into broader discrete intervals reduces individual distinguishability ($k = 1$) while retaining analytical use of the microdata. This experiment provides a focused baseline for this mitigation family, leaving optimized disclosure-control designs to future work.

On the complete-score 2023 cohort ($N = 2,678,264$), we compare the original one-decimal representation, integer rounding, and fixed-width bins of 5, 10, 25, 50, 100, 150, 200, 250, and 300 points. Each transformation is applied separately to the five performance scores. Integer rounding uses the implementation's standard round-to-even rule. For a bin of width B , score s is represented by its bin center, $B\lfloor s/B \rfloor + B/2$. The reduced average is then recomputed from the five transformed scores and rounded to two decimals.

Privacy is measured as the proportion of records in singleton equivalence classes ($k = 1$), but the attack key depends on adversary capability. The Neighbor recomputes the average from the released reduced scores and combines it with gender, municipality of examination, and age range. The Specialist and Profiler instead hold an original-precision five-score tuple from an auxiliary source and apply the release's transformation to their own tuple before exact lookup; the Profiler additionally retains gender, municipality, and age range, which score-only generalization does not alter. We repeat this privacy calculation within the score-based population subsets and municipal-density strata defined in Section 5.2 to test whether the population rate masks heterogeneous protection.

We assess analytical utility on the same 2023 cohort using four outcomes. Spearman correlation measures preservation of candidate

ordering, and mean absolute error (MAE) measures per-candidate distortion of the average score. Following the educational-equity uses summarized by Franco et al. (2024), we also measure the signed change in the private-minus-public school gap and the higher-minus-lower family-income gap. These metrics cover ranking, score fidelity, and two prominent inequality comparisons; they do not represent every scientific use of ENEM microdata. Accordingly, the experiment characterizes the tested privacy-utility profile and does not establish a universal acceptability threshold or show that other disclosure-control mechanisms cannot perform better.

Recalculability. Score-only defenses must also be interpreted in relation to the other released artifacts. ENEM objective-test scores use a three-parameter IRT model, and public release packages combine candidate-level item responses with calibrated item parameters; under INEP's documented procedure, those inputs permit independent estimation of the four objective-test proficiencies (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2021, pp. 15–21). The essay is different: its competency scores are assigned by human evaluators and cannot be derived through the same IRT calculation. Where item-level responses and corresponding item parameters remain available, rounding or binning only the published objective scores may be bypassed by estimating finer-resolution values. Release-specific field availability is listed in the supplement's recalculability family.

5.5. Implementation and computational setup

All analyses were implemented in Python using chunked data processing and precomputed lookup indices. The core runs were executed on a consumer laptop with a 12th-generation Intel Core i5-1240P processor and 16 GB of RAM. Across separately executed runs, the recorded wall-clock times totaled approximately 6.87 h: 2.86 h for the ENEM k -anonymity audit across 15 editions and 57.7 million complete-score records, 13.1 min for Vector A, and 3.79 h for the Vector C baseline and three deduplication-key variants. These measurements show that the evaluated threat models do not require high-performance computing.

6. Results

Results follow the sequence established in Section 5: longitudinal distinctiveness and its 2023 anatomy, empirical linkage through Vectors A–C, the privacy-utility profile of score generalization, and the attributes exposed.

6.1. Longitudinal audit: The persistence of risk

The audit covers the currently hosted Phase 2 packages for all fifteen ENEM editions (2009–2023) and 57,708,475 complete-score candidate records (67.0% of 86,110,930 enrollment records). Distinctiveness persists across all fifteen editions (Fig. 2). The Neighbor adversary, which uses the average score, gender, municipality of examination, and age range, yields a mean uniqueness ($k = 1$) of 92.62%, ranging from 89.85% in 2015 to 94.84% in 2020. The Specialist's exact five-score tuple is still more distinctive: its edition-level mean is 99.9985%, and adding the Profiler's demographic context raises the mean to 100.0000%. The tuple's near uniqueness is a longitudinal property rather than a feature of one cohort.

6.2. Anatomy of the linkage: The 2023 snapshot

Among $N = 2,678,264$ complete-score candidates in 2023, there were $M = 2,678,235$ distinct tuples and 2,678,225 candidates in singleton classes. Table 5 shows that average score and gender alone provide little identifying power, whereas geographic knowledge produces the main increase among average-based adversaries: uniqueness reaches

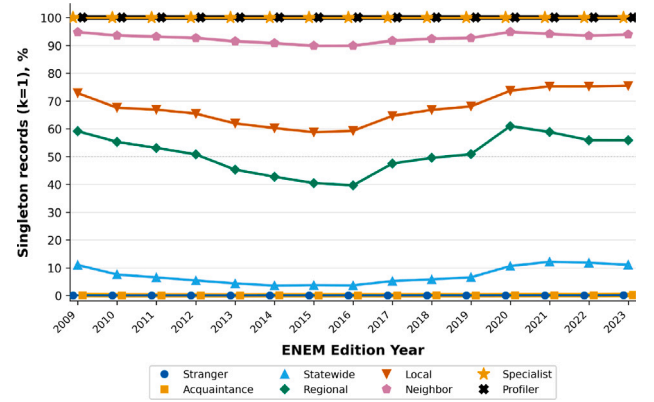


Fig. 2. Longitudinal uniqueness ($k = 1$) by adversary model across the currently hosted Phase 2 packages for fifteen ENEM editions (2009–2023).

Table 5

Record individualization as quasi-identifiers accumulate (ENEM 2023).

Adversary	Available QIs	Unique ($k=1$)
Stranger	Average score	0.11%
Acquaintance	+ Gender	0.21%
Statewide	+ Gender + State	11.07%
Regional	+ Gender + State + Age range	55.86%
Local	+ Gender + Municipality of exam	75.51%
Neighbor	+ Gender + Municipality of exam + Age range	93.94%
Specialist	5-Score tuple	99.9985%
Profiler	5-Score tuple + Gender + Municipality of exam + Age range	100.00%

75.51% with municipality and 93.94% when age range is added. The exact score tuple alone reaches 99.9985%, with demographics resolving the remaining collisions.

Fig. 3 shows that population averages conceal greater exposure in the sparse upper tail. Neighbor uniqueness rises from 93.94% overall to 98.51% in the top 1% and 99.78% in the top 0.1%. The gradient is stronger for weaker adversaries: within the top 0.1%, average score alone identifies 18.41%, adding gender identifies 36.49%, and adding state raises uniqueness to 90.42%, compared with 0.11%, 0.21%, and 11.07% in the full cohort.

The municipality scatter (Fig. 4) reveals a complementary density effect. In São Paulo ($n = 105,986$), the Local key identifies only 9.62%, but age range raises Neighbor uniqueness to 57.90%. In Pelotas ($n = 6640$), municipality already identifies 81.20%, so age contributes a smaller 16.27 percentage points. In Santa Rosa do Purus ($n = 16$), the Local key already identifies every candidate and age adds nothing. The supplement's score-distinctiveness family provides the complete annual series and the 2023 performance-stratum and municipality aggregates.

6.3. Empirical validation of attack vectors

The audit establishes that score-based keys distinguish records; the attack vectors test whether independent public sources associate those keys with identities. They progress from reconstruction of an institution-specific score and conditional auxiliary attributes (A), through an identity-associated tuple that circulates outside ENEM (B), to exact linkage with governmental open data (C).

6.3.1. Vector A: Institutional linkage

For the 2024 UFRGS cycle ($n = 9455$ across 100 programs), uniqueness remains low without geographic knowledge: 0.30% for the Stranger and 0.60% for the Acquaintance under the full pipeline. Geographic knowledge changes the result sharply (Fig. 5). Statewide

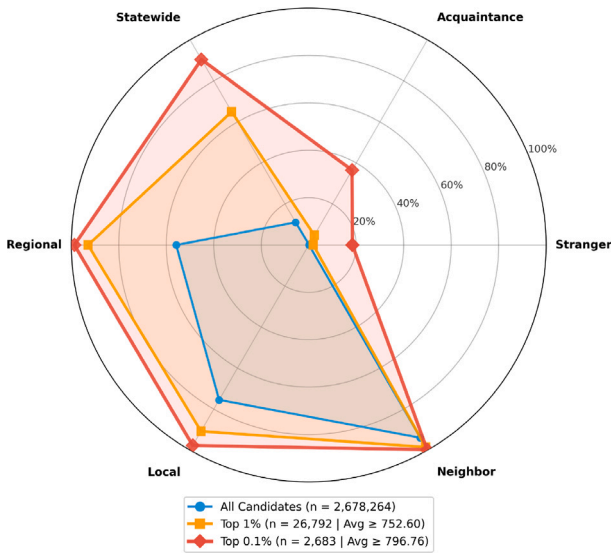


Fig. 3. Uniqueness ($k = 1$) by performance stratum for the six average-based adversaries shown, ENEM 2023.

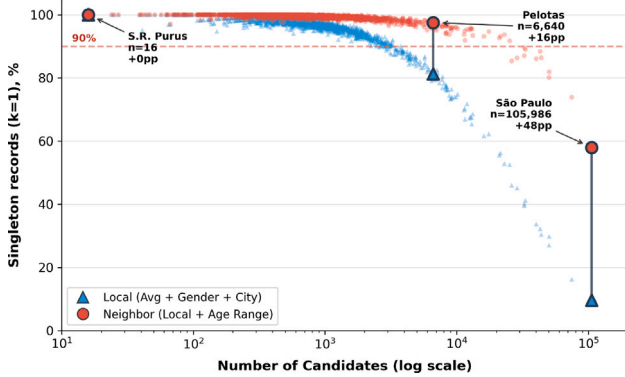


Fig. 4. Local and Neighbor uniqueness by municipality candidate-pool size, ENEM 2023. Highlighted gaps show the marginal contribution of age range after average score, gender, and municipality are known.

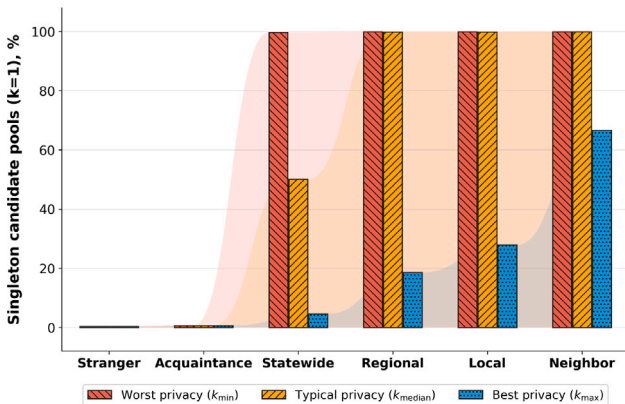


Fig. 5. Vector A uniqueness ($k = 1$) across UFRGS adversaries ($n = 9455$). For scenarios with unobserved auxiliary QIs, $k_{\min}$, k_{median} , and $k_{\max}$ summarize the worst, typical, and best privacy outcomes across the evaluated geographic and age values.

uniqueness spans 99.68% under $k_{\min}$, 50.09% under the typical k_{median} , and 4.59% under $k_{\max}$. With municipality, the corresponding Local

Table 6

Geographic saturation of inferred filters — UFRGS SISU 2024, $n = 9455$. All adversaries operate with the full Vector A inference pipeline (gender and quota filtering) under Typical geographic scenarios; the *Excl. Inference* column reports the counterfactual uniqueness when those filters are removed, holding geographic knowledge fixed.

Adversary	Excl. inference	Full pipeline	Δ
Stranger	0.27% (26)	0.30% (28)	+0.03 pp
Acquaintance	0.27% (26)	0.60% (57)	+0.33 pp
Statewide	13.81% (1306)	50.09% (4736)	+36.28 pp
Regional	99.83% (9439)	99.82% (9438)	−0.01 pp
Local	99.86% (9442)	99.85% (9441)	−0.01 pp
Neighbor	99.89% (9445)	99.90% (9446)	+0.01 pp

rates are 99.88%, 99.85%, and 27.87%; adding age range raises the Neighbor rates to 99.90%, 99.90%, and 66.69%. Because geography and age are scenario assumptions rather than attributes observed in the admission list, these are conditional bounds, not verified nominal associations.

Table 6 compares the main run with a counterfactual run excluding gender and quota inference. The largest difference occurs at the Statewide level, where typical uniqueness rises from 13.81% to 50.09% (+36.28 pp). By contrast, Regional, Local, and Neighbor change by only one candidate each (−0.01, −0.01, and +0.01 pp⁴), showing that their high rates arise primarily from the score pivot and assumed geography.

The aggregate results are invariant across round-to-even, round-half-up, and truncation, and the tie audit found no ambiguous rounded-tuple groups. Adaptive fallback relaxed a quota heuristic for four candidates (0.04%); those cases retain a conservative k bound but cannot support attribution of the filtered demographic profile. Robustness outputs and the full filter matrix are provided in the supplement's vector-a-institutional family.

6.3.2. Vector B: Self-disclosure linkage

Vector B illustrates how the linkage key can circulate through candidate action, bridging Vector A's reconstructed institutional score and Vector C's consolidated governmental source. Our documentary review found public posts and screenshots on X, Instagram, TikTok, and Reddit associating ENEM scores with identifiable profiles. Where a post provides an identifiable profile, a complete five-score tuple, and its ENEM edition, a singleton lookup could associate that profile with attributes in the matching microdata record. We also inspected SISU-simulator interfaces that request scores in an account or contact context, including Descomplica, ProENEM, and LazuDedu.⁵ These platforms typically request the five scores together with account or contact data (email, social login, or phone). If an operator could access the complete tuple together with that identity context and the corresponding ENEM edition, those inputs would permit an exact lookup in the relevant microdata, succeeding when it returns a single record ($k = 1$). Under those conditions, no geographic knowledge, demographic inference, or score reconstruction is required. This result establishes the conditional mechanism and documents concrete routes through which its inputs may circulate, not their population prevalence: we did not retain a disclosure corpus, access platform records, or estimate how often complete tuples circulate.

⁴ For each candidate, the typical bound is the median across locations yielding at least one match. Filtering can remove a location from this set, so the median may change in either direction.

⁵ Simulator interfaces inspected (accessed July 15, 2026): Descomplica (<https://descomplica.com.br/sisu/simulacao/>), ProENEM (<https://sisu.proenem.com.br/>), LazuDedu (<https://lazedu.com/simulador-sisu/>).

Table 7

Governmental linkage funnel (2010–2023). The upper block reduces raw SISU rows to candidate-years; the middle block reports exact five-score linkage outcomes; and the lower block converts unique candidate-year matches into longitudinal estimates of distinct individuals.

Stage	Count	Δ
SISU rows to candidate-years		
Raw ingestion (Chamada Regular)	88,050,294	–
After removing digit-bearing names	88,050,243	–51
After removing invalid scores (NaN 38,118; zero 68,914)	87,943,211	–107,032
Valid SISU rows	87,943,211	–
Valid candidate-years after annual deduplication	26,612,533	–61,330,678
Deterministic join vs. ENEM microdata		
Unique matches ($k = 1$)	26,589,382	
Ambiguous matches ($k > 1$)	77	
Not found ($k = 0$)	23,074	
Unique-match rate ($k = 1$)	99.91%	
Unique-person accounting (cross-year)		
Cross-year dedup (full key, pre-canonicalization)	14,797,510	–11,791,872
Canonical name map	14,688,801	–108,709
Estimated individuals (post-canonicalization)	14,688,801	

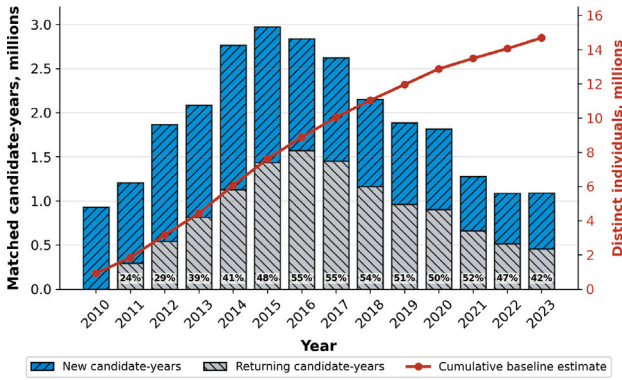


Fig. 6. Governmental linkage across 14 annual SISU cohorts (2010–2023). Stacked bars decompose uniquely matched candidate-years into first appearances (*New*) and applicants linked in an earlier cohort (*Returning*); annotations give the returning share. The red line traces the cumulative estimate under the full deduplication key and canonical name map, reaching 14.69M distinct individuals in 2023.

6.3.3. Vector C: Governmental linkage

Vector C links two complementary federal releases through the exact five-score tuple. We report two distinct quantities. A *candidate-year* is the unit of linkage because each SISU cohort is matched to its corresponding ENEM edition; a person who applies in multiple years can therefore contribute multiple matches. A *distinct individual* is instead a longitudinal estimate obtained by deduplicating those candidate-year matches across cohorts. Table 7 keeps the two units separate from ingestion through longitudinal accounting.

The same concern for accurate counting motivates a restriction on the linkage corpus. The source-selection audit supports limiting it to regular-admission files. Of the 14,649,945 normalized names obtained by summing the annual unique waitlist counts, only three were absent from the corresponding *Chamada Regular* corpus.

The linkage outcome. The exact five-score join uniquely matched 26,589,382 of 26,612,533 valid candidate-years (99.91%). Only 77 candidate-years returned multiple ENEM records ($k > 1$), and 23,074 returned no match ($k = 0$), plausibly reflecting records or low-frequency categories omitted from the public microdata under INEP's

Table 8

Sensitivity of the unique-person count to the deduplication key (2010–2023). All variants share the cross-year canonical name map, isolating the contribution of each key attribute.

Variant	Key	Unique	Δ vs. baseline
full (baseline)	Name Year CPF ₄₋₈	14,688,801	–
no_year	Name CPF ₄₋₈	14,686,159	–2,642
no_cpf	Name Year	14,210,958	–477,843
name_only	Name	12,654,034	–2,034,767

disclosure controls (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2024, p. 7).⁶ Cross-year deduplication of the unique matches yields an estimate of 14,688,801 distinct individuals. The gap between the two counts arises from applicants returning in later SISU cohorts (Fig. 6).

Sensitivity of the unique-person count. The 14.69M estimate depends on how recurring applicants are recognized, so we re-executed the pipeline under four deduplication keys (Table 8). The CPF fragment is the decisive discriminator. While it is present, the birth year is nearly redundant (–2642; –0.02%). Without the fragment, however, birth year has a much larger effect: name plus year yields 14.21M, whereas name alone yields 12.65M.

The remaining uncertainty concerns person counting, not the observed candidate-year linkage outcomes. Legal name changes or unmodeled spelling variants can split one person across keys and inflate the estimate; homonyms or erroneous similarity merges can instead collapse different people and reduce it. Their residual magnitudes were not empirically estimated. As a result, we treat 14,688,801 as an estimate under the specified full key and canonicalization rules, not as an exact civil census. Exact similarity rules, canonicalization effects, score-format handling, affected CPF-mask files, annual series, extended sensitivities, telemetry, and synthetic tests are documented in the supplement's *vector-c-governmental* family.

6.4. Evaluating defenses: Statistical Disclosure Control (SDC)

Score generalization reduced uniqueness, but not uniformly across adversaries or candidates (Table 9). For the average-based Neighbor,

⁶ The pipeline did not classify the unmatched cases by release-side cause; documented suppression is hence a plausible contributor rather than an attribution for every $k = 0$ outcome.

Table 9

Selected privacy–utility results for score generalization, ENEM 2023 ($N = 2,678,264$). Privacy columns report uniqueness ($k = 1$); utility columns report Spearman rank correlation, average-score MAE in points, and signed school- and income-gap distortions (generalized – original) in points.

Resolution	Privacy ($k = 1$, %)			Utility			
	Neigh.	Spec.	Prof.	Spear.	MAE	Sch.	Inc.
Original	93.94	>99.99	100.0	1.000	0.0	0.0	0.0
Int. round.	72.80	99.98	100.0	1.000	0.1	0.0	0.0
Bin 50	12.88	0.60	84.55	0.998	5.3	−0.2	−0.1
Bin 100	7.53	0.05	47.39	0.990	10.6	−0.3	−0.2
Bin 250	3.32	0.00	13.38	0.954	29.5	+3.9	+3.2
Bin 300	2.83	0.00	7.76	0.905	37.8	+5.3	+7.3

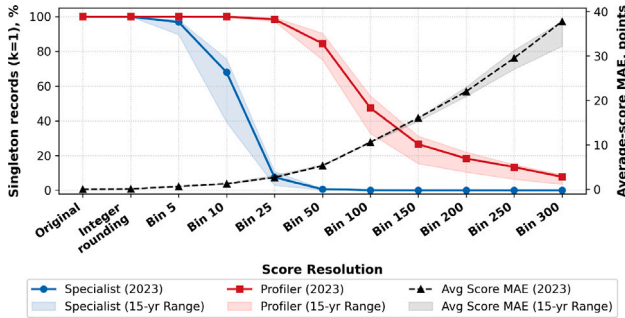


Fig. 7. Privacy–utility profile of score generalization. Specialist and Profiler uniqueness ($k = 1$) and average-score MAE are shown for ENEM 2023, with the range across the 15 ENEM editions (2009–2023) shaded for each curve. Binning rapidly suppresses the score-only Specialist, whereas the Profiler declines only under coarser resolutions.

integer rounding lowered $k = 1$ from 93.94% to 72.80%, and 50-point bins lowered it to 12.88%. Yet this population average concealed unequal protection: at bin 50, 53.52% of the top 0.1% of scorers and 33.02% of candidates in municipalities with fewer than 1000 examinees remained in singleton classes, compared with 0.38% in the largest municipalities.

The tuple adversaries reveal why the same transformation protects unevenly (Fig. 7). Binning defeats the Specialist early, because it acts directly on the only information that adversary holds: the five-score tuple. The Profiler resists longer, since gender, municipality, and age range survive the binning and still separate candidates whose tuples have merged. The rising MAE curve marks the utility cost of coarser bins, and the shaded 15-edition envelopes show that this contrast is not specific to 2023. Generalizing scores alone protects against the score-only threat model well before it protects against linkage augmented by unchanged quasi-identifiers. These results measure record individualization through $k = 1$; they do not quantify attribute disclosure within the larger classes that binning creates.

The utility side of this trade-off runs along the same resolution axis. Across the finer resolutions, candidate ordering, average-score fidelity, and the two inequality gaps remain close to their original values. By the time coarsening substantially suppresses the Profiler, however, score error has accumulated and the group-gap estimates have begun to shift (Table 9). No tested resolution eliminated Profiler singleton records while keeping all four evaluated utility outcomes close to their original values. This is an empirical profile of fixed score generalization, not an acceptability threshold or evidence that other disclosure-control designs cannot perform better. The supplement's score-generalization family provides the complete resolution grid and annual and stratified aggregates.

Table 10

SISU attributes connected to ENEM records by Vector C. Fields were verified across the real 2010–2023 SISU releases; availability varies across mapped release eras.

Category	Attributes
Civil identity	Full name; masked CPF (six central digits in the standard mask, with a shifted fragment in affected files); gender; and birth year (2010–2015, 2023) or full date (2016–2022)
Residence	Candidate municipality and state
Program choice	Institution, campus, course, degree, shift, and first/second option
Performance	Five raw ENEM scores, program weights, weighted scores, candidate score, and cutoff score where available
Admission outcome	Classification rank, approval, enrollment, and selection-stage fields
Admission modality	Competition/quota modality and description, including public-school, income, racial, and disability criteria where encoded.

Table 11

Selected ENEM attributes made attributable after linkage, as documented in official release headers and data dictionaries. Availability and missingness vary by edition and candidate.

Category	Attributes
Demographics and place	Age range, gender, marital status, race/color, nationality, and municipality/state of examination
Family and inequality	Education and occupation of parents or guardians, household size, family income, and employment of a domestic worker in the household
Housing and assets	Bathrooms, bedrooms, vehicles, appliances, telephones, computer, and internet access
Performance	Four objective-test scores, five essay-competency scores, and total essay score
School context	School type, municipality/state, administrative dependency, urban/rural location, and operating status
Exam behavior	Foreign-language choice and item-level responses (marked responses to the four objective tests)

6.5. Synthesis: The human impact

The principal human impact is the conversion of two separately released federal datasets into an identity-linked dossier. Through the exact score tuple, Vector C combines SISU civil identity, program choices, admission outcomes, and quota modality with ENEM socioeconomic, school, and performance attributes—including family income, housing conditions, household assets, and the complete five-score profile (Tables 10 and 11). The analysis estimates 14,688,801 distinct individuals with linked records; specific fields vary by edition and candidate.

6.5.1. Accessibility-related exposure

A distinct exposure arises from exam-book codes. Booklets adapted for accessibility carry category-specific codes that act as proxies for accessibility-related attributes, including Braille, enlarged and super-enlarged print, reader assistance, Libras, and generic specialized care (Table 12). Some of these are proxies for specific disabilities: Braille booklet codes for visual impairment and Libras booklet codes for deafness, categories the LGPD treats as sensitive health data (Brazil, 2018, Art. 5, item II). A code does not identify a candidate on its own, as the smallest observed equivalence class still contained 103 candidates; after an identity–score linkage, however, it becomes an immediately readable attribute of the linked record. Across the currently hosted Phase 2 packages for 2009–2023, this amounts to 26,295 candidate-edition exposures, or 0.031% of 86.1 million enrollment records.

This channel persists under the disjoint Phase 3 architecture. Although the 2024 RESULTS file no longer carries the socioeconomic

Table 12

Accessibility-related exposure through exam-book codes. Counts are distinct candidates within each year and candidate-edition exposures when summed across years. The categories partition the exposed set within every edition; dashes indicate no observed exposure.

Year	Reader	Libras	Braille	Enlarged	Super-enl.	Spec. care	Total
2009–2023 (Phase 2)							
2009	131	–	–	–	–	–	131
2010	–	–	–	–	–	–	0
2011	–	–	–	–	–	–	0
2012	639	–	–	–	–	–	639
2013	977	–	–	–	–	–	977
2014	–	–	–	–	–	1,125	1,125
2015	–	–	–	–	–	3,135	3,135
2016	–	–	–	–	–	1,073	1,073
2017	591	2,545	–	–	–	–	3,136
2018	570	2,197	–	–	–	–	2,767
2019	533	2,222	–	–	–	–	2,755
2020	387	1,234	–	–	–	–	1,621
2021	330	930	–	–	–	–	1,260
2022	353	1,025	–	–	–	–	1,378
2023	716	1,287	125	3486	684	–	6,298
Total	5227	11,440	125	3486	684	5333	26,295
2024 (Phase 3)							
2024	1433	1,304	–	3528	750	–	7,015
Total	1433	1,304	0	3528	750	0	7,015

questionnaire, it retains scores and adapted-booklet codes: 7015 candidates bear codes across four accessibility categories, corresponding to 0.162% of 4.3 million enrollment records. The longitudinal sum is therefore 33,310 candidate-edition exposures, not 33,310 unique people; a candidate appearing in multiple editions may be counted more than once. Complete code mappings and annual aggregates are provided in the supplement's schema catalog and accessibility-proxies family, respectively.

7. Discussion

The results identify a disclosure problem produced by the composition of data releases. Across fifteen ENEM editions, multidimensional scores remained highly distinctive; Vectors A–C then showed how score-bearing sources progressively reduce the effort required to associate those records with identities. Independently published federal data turned this statistical property into deterministic population-scale linkage. The policy question consequently extends beyond the suppression of selected demographic fields to the full release environment in which scores, identities, response data, and sensitive attributes coexist.

7.1. Inter-agency governance and the complementary-release trap

INEP's disclosure controls and the publication of admission data by MEC/SESU and public institutions respond to legitimate but separately administered objectives: statistical confidentiality, educational transparency, and auditable selection. The resulting releases nevertheless interact. ENEM microdata retain the same precise scores that SISU records and institutional lists associate with civil identities, creating the complementary-release condition described by Sweeney (Sweeney, 2002). Under the tested architecture, removing School Code or residential identifiers reduced some demographic linkage paths but did not remove the score pivot. The longitudinal persistence of tuple distinctiveness and the exact governmental join therefore show that demographic or geographic suppression alone does not address this threat model.

The prior INEP-commissioned assessment concluded, from quantitative experiments using demographic and institutional auxiliary information, that the de-identification and pseudonymization techniques then used by INEP were insufficient to mitigate re-identification and sensitive-attribute inference risks (Alvim et al., 2020, p. 7). Our results establish the ENEM-specific mechanism that warning left open, and

the three vectors clarify how auxiliary sources reshape the adversary model. Vector A is conditional because the adversary must reproduce an institution's published program-weighted mean from ENEM records, infer demographic filters, and know geography and age. Vectors B and C instead use the highly distinctive exact five-score tuple. Vector B depends on individual or mediated disclosure, whereas Vector C obtains both the scores and civil identity directly from a governmental release. Richer complementary sources therefore remove adversarial assumptions progressively.

The experiments also provide technical evidence relevant to the LGPD's reasonable-effort criterion (Brazil, 2018, Art. 12). This evidence changes the technical basis of the threat model contested in the Public Civil Action. The MPF argued that the assumed re-identification scenario fell outside what could reasonably justify the restrictions because it depended on participants maintaining open social-media profiles containing birthplace, residence, age, and school information and on manual, individualized analysis of those profiles (Ministério Público Federal, 2022, pp. 15–16); Vector C instead performs an exact population-scale join between federal releases without demographic inference. The sources were publicly obtainable, and the evaluated core runs used a consumer laptop, totaling approximately 6.87 h across separately executed audits and sensitivity runs. These facts characterize availability, computational effort, and operational feasibility for the tested attacks; the legal assessment of whether that evidence satisfies Article 12 belongs to the competent authorities.

7.2. Privacy–utility profile and recalculability

The judicial sequence surrounding ENEM makes the engineering trade-off concrete. Decision A ordered anonymization techniques that preserve analytical utility (Justiça Federal da 1ª Região, 2025, Sec. 5 and dispositive item (b)), whereas Decision B recognized substantial compliance based on INEP's current disclosure controls, including the separated 2024 architecture (Justiça Federal da 1ª Região, 2026, Secs. 1 and 2.3; dispositive). Separating participant attributes from results blocks the socioeconomic linkage, but it also prevents analyses that require performance and demographic or socioeconomic variables in the same analytical unit. Conversely, restoring a monolithic release with the original scores would restore the linkage surface demonstrated here. The current options occupy two unsatisfactory poles: socioeconomically unlinkable but relationally constrained, or relationally useful but demonstrably linkable. The relevant question is which release

mechanisms can provide specified analytical uses under an explicit adversary model.

Our SDC experiment supplies a bounded answer for fixed score generalization. Coarsening reduced Profiler individualization, but even the coarsest resolution tested left 7.76% of records singleton, and all four utility outcomes showed their largest departures from the original representation at that resolution. This tension parallels Daries et al.'s finding that de-identification of MOOC records, guided by FERPA requirements, materially degraded analytical utility (Daries et al., 2014). For the Neighbor, protection was also heterogeneous: under 50-point bins, 53.52% of the top 0.1% and 33.02% of candidates in municipalities with fewer than 1000 examinees remained in singleton classes. Together, these findings show that fixed generalization reduced measured individualization without eliminating it and that protection remained uneven across candidate groups.

Score-only transformations leave a recalculability problem: every audited public ENEM microdata package releases 180 item-level responses per candidate alongside calibrated item parameters, which permit independent estimation of the four objective-test proficiencies under INEP's documented procedure (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2021, pp. 15–21). Rounding, binning, or perturbing only the published objective scores may therefore be bypassed through re-estimation; the essay is not equivalently recalculable. Release design must address scores, responses, parameters, and auxiliary data jointly. Controlled access, query systems, jointly designed perturbation, differential privacy, and synthetic data remain possible, but require explicit analytical targets and population-scale validation. Decision A imposed a 180-day implementation deadline (Justiça Federal da 1ª Região, 2025, dispositive item (c)), a compressed window for designing and validating a new population-scale release mechanism. U.S. Census modernization illustrates the sustained engineering involved in deploying differential privacy (Abowd et al., 2020). In its motions for clarification, INEP presented its current controls, including its pre-existing 2024 separation architecture, as compliance with Decision A. Decision B recognized substantial compliance and revoked the deadline (Justiça Federal da 1ª Região, 2026, Secs. 1 and 2.3; dispositive). This avoided a rushed redesign that might have expanded the demonstrated linkage surface, but left an architecture that sidesteps rather than resolves the release-design problem, sacrificing relational utility while retaining the residual exposure documented in Table 12. Population-scale anonymization mandates therefore require sufficient implementation time and explicit privacy–utility criteria for a validated replacement.

7.3. Residual vulnerability, legacy privacy debt, and potential harm

The 2024 separation blocks the historical socioeconomic linkage within the released architecture. Accessibility-related booklet codes nevertheless remain beside scores in the RESULTS file. In 2024, 7015 candidate-edition records bear such codes. Separation alone does not guarantee protection for sensitive proxies retained beside a distinctive key.

Historical Phase 1 releases create a different problem. They placed precise scores and detailed participant attributes in the same candidate-level record, including edition-specific specialized-assistance indicators concerning autism, pregnancy, prison-unit status, and social-name use, among other personal circumstances (Willrich, 2022, Sec. 2.3). Later disclosure controls cannot recall copies already distributed. In 2018, the newspaper *Folha de S. Paulo* reported that it had analyzed anomalous ENEM answer patterns and cross-referenced the microdata with information on university entrants, identifying a candidate associated with a suspicious response pattern (de S. Paulo, 2018). That civic-audit use demonstrates that microdata-based identity linkage was feasible for a non-state actor before the attacks evaluated here. The legacy privacy debt lies in the retained copies of those releases: after a successful identity–score linkage, the attributes present in the corresponding

edition become attributable to the linked person. This debt is release-specific because not every field existed in every year and not every linked record contains every attribute. However, the persistence of score distinctiveness across fifteen editions makes the attack surface durable.

The study did not execute downstream abuse, so harms beyond linkage remain plausible consequences that were not observed in our experiments. The joined data nevertheless support a concrete downstream threat model. An attacker impersonating an admissions office could incorporate a target's course, institution, ranking, residence, or family context into spear-phishing messages. Alongside the civil name in the same SISU record, a masked CPF fragment and, depending on the release era, a birth year or full birth date reduce the candidate space for identity resolution and could support subsequent account-recovery or fraud attempts. A mask-displacement error adds a release-composition risk. Affected files expose CPF positions 3–8, while the standard mask exposes positions 4–9. If the same candidate appears under both variants, combining the fragments can reveal positions 3–9. The join also enables attribute-based target lists. An adversary could, for instance, select Medicine applicants at federal universities whose ENEM questionnaires report family income above 20 minimum wages and high-value household assets, supporting financial targeting or discriminatory profiling in credit scoring, employment screening, or insurance underwriting. This linkage uses Open-Source Intelligence (OSINT), drawing on publicly downloadable sources rather than breached systems.

7.4. Beyond ENEM: Generality and its preconditions

Although ENEM provides the empirical case, the underlying governance problem is not uniquely Brazilian. Large-scale assessments (e.g., SAT, NAEP, GRE, and PISA) use IRT-family models to estimate proficiency (College Board, 2024; National Center for Education Statistics, 2024; Wendler and Bridgeman, 2014; OECD, 2024). IRT use alone does not reproduce the vulnerability: systems differ in score precision, dimensionality, sampling, microdata availability, response-data access, and admission practices. Nevertheless, IRT-family systems can produce granular score combinations with the linkage-relevant property observed here: rarity within a released population. Coarser or lower-dimensional reporting may reduce, but not necessarily eliminate, this distinctiveness; whether it persists must be tested empirically in each system. Identity linkage additionally requires an obtainable source associating the same values, or a derivable equivalent, with identities; absent one, distinctiveness can establish individualization but not identity linkage, and future linkage remains conditional on such a source becoming available.

INEP's impact assessment lists performance scores but does not assess them as linkage keys, confining mitigation to geographic and demographic suppression (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022a, Secs. 3.1.10–3.1.28, 3.2.1, and 7). This illustrates a broader custodial gap. Scores should therefore be tested as cross-release pivots, particularly in sparse performance tails and small geographic populations. The release ecosystem, not a single dataset or agency, is the unit of governance. Cross-custodian threat modeling, linkage testing, and explicit privacy–utility objectives should precede publication whenever precise performance data coexist with identifiable auxiliary records.

8. Ethical considerations and coordinated disclosure

Under Resolution CNS No. 510/2016 of the Brazilian National Health Council (Art. 1, sole paragraph, item II, read with Art. 2, item VI) (Conselho Nacional de Saúde, 2016), research based exclusively on publicly accessible information is exempt from registration and evaluation by the Brazilian Research Ethics Committee system (CEP/CONEP). Accordingly, no ethics-board review or waiver was required. The study

used datasets made publicly available by governmental agencies and a public higher education institution, together with information voluntarily posted on unrestricted online platforms. The governmental and institutional releases are subject to the Brazilian Access to Information Law (Law No. 12.527/2011) (Brazil, 2011). The study involved no interaction with human participants and did not solicit or collect information directly from individuals.

The linkage analyses were performed solely to assess disclosure risk. No nominal cross-dataset linkage dataset was retained, and no individual association is reported. The supplement contains only aggregate, declarative, and synthetic materials. Confidential access is limited to code inspection; restricted data and record-level outputs are not shared.

We initiated Coordinated Vulnerability Disclosure with INEP on February 27, 2026, before publicly disseminating the technical details reported here. We participated in technical meetings, presented the methodology and findings, and provided a written summary of the mechanism, root cause, and potential impact. A 90-day embargo was established to support assessment and remediation. During the embargo, INEP engaged the Ministry of Education, which administers the SISU releases, and the meetings covered revision of the Data Protection Impact Assessment and a secure architecture for future microdata publication. The research team was invited to contribute to this work through a formal working group.

At the end of the embargo in late May 2026, the historical ENEM microdata and nominal SISU reports used by the study remained publicly accessible in their original formats. Because previously distributed files cannot be recalled from offline copies, the historical linkage risk persists even if official repositories are later modified. We therefore disclose the scientific method and aggregate evidence needed to evaluate the claims while withholding artifacts that would lower the barrier to record-level exploitation.

9. Conclusion

This work demonstrates that the scale of ENEM microdata did not create safety in numbers. Across fifteen editions and 57.7 million complete-score records, precise academic performance scores are not merely analytical outcomes but high-entropy quasi-identifiers. Under the Specialist's exact five-score model, nearly all complete-score records belonged to singleton equivalence classes ($k = 1$). Vectors A–C then established progressively lower-friction routes from performance data to identity, making the attributes in linked records attributable to individuals. The strongest case, an exact five-score join between federal releases, uniquely matched 26.59 million of 26.61 million candidate-year records in the federal admission system (99.91%). Cross-year canonicalization yielded an estimate of 14.69 million distinct individuals with linked records. Scale did not provide a protective crowd; combined with precise scores and complementary releases, it gave the linkage surface a reach measured in millions of records and people.

These attack vectors used publicly obtainable sources and consumer hardware, providing technical evidence about the availability and operational effort relevant to an assessment under the LGPD's reasonable-effort criterion. The SDC experiment likewise showed that fixed score generalization reduced but did not eliminate Profiler individualization, with the coarsest tested resolution producing the largest departures across the four evaluated utility outcomes. Individualization remained higher among top performers and candidates in small municipalities.

No mechanism evaluated in this study fully resolves the disclosure problem, and several additional approaches merit systematic evaluation. Separating socioeconomic attributes from scores blocks that linkage channel in the 2024 ENEM microdata release but does not eliminate disclosure risk. Score-linked exam-book codes leave a conditional exposure in that edition, while the previously distributed 2009–2023 files remain available and create a legacy risk. Future release design should begin with explicit analytical uses and adversary models and

jointly assess scores, item-level responses, item parameters, and predictable auxiliary sources. Candidate approaches include controlled access, query systems, jointly designed perturbation, differential privacy, and synthetic data. Our stratified results further motivate evaluating dynamic aggregation or thresholding designed to strengthen protection for high-performing candidates and small municipal cohorts. All require population-scale validation of privacy and utility. Investigating these potentially higher-utility alternatives is preferable to treating the relationally constrained 2024 separation as a permanent endpoint. Any replacement must satisfy explicit privacy criteria under the demonstrated linkage threats.

Future work will pursue two directions opened by these findings. First, we plan to design and empirically validate disclosure-control mechanisms tailored to high-entropy performance representations, including optimization-based generalization and synthetic-data generation, evaluating both privacy and analytical utility at population scale. Second, we intend to apply the disclosure-assessment methodology to other public repositories of multidimensional performance data, from professional-certification systems to healthcare outcome records, to test whether the score-as-linkage-key mechanism generalizes across domains.

The threat model demonstrated here reaches beyond the Brazilian case. High-resolution multidimensional performance variables should be treated as quasi-identifiers during disclosure-risk assessment, not as purely analytical attributes. Such variables warrant linkage testing whenever obtainable sources can associate the same values, or derivable equivalents, with identities. For Brazil specifically, the trajectory from Decision A's order to restore transparency to Decision B's recognition of substantial compliance through a fragmented architecture illustrates two risks of reactive governance. Hasty restoration mandates can sacrifice privacy, whereas treating formal compliance as a settled endpoint can permanently forgo the relational analytical value of open educational data. We therefore urge regulatory and judicial bodies to weigh both failure modes explicitly before mandating renewed disclosure or accepting fragmentation.

CRedit authorship contribution statement

Henrique Lindemann: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Eder J. Scheid:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation. **Lisandro Z. Granville:** Writing – review & editing, Validation, Supervision, Resources. **Muriel F. Franco:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used Claude and ChatGPT to assist with writing, paraphrasing, and structural revision of portions of the text. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The primary datasets analyzed in this study are publicly available: the ENEM microdata from the INEP Open Data Portal <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/enem> and the MEC/SISU Open Data from the MEC Open Data Portal <https://dadosabertos.mec.gov.br/sisu>, both accessed on July 15, 2026. To support scientific verification, we provide a supplementary package <https://github.com/ComputerNetworks-UFRGS/enem-microdata-privacy> containing sanitized aggregate results underlying the reported tables and figures, scripts that verify the claims and regenerate the corresponding tables and figures from those aggregates, claim-facing schema documentation for the official ENEM and MEC/SISU releases and exam-book codes, pseudocode for each linkage pipeline, and synthetic-data examples illustrating the matching logic. Evidence-family labels in the text match the supplement's directory names; each family README gives its scope and points to the relevant files. The operational linkage code is not publicly released, as its distribution would materially lower the barrier to the attacks documented in this work. Reviewers and competent regulatory and data protection authorities may request confidential, code-only inspection of this material through the corresponding author for verification purposes.

References

- Abowd, J.M., Benedetto, G.L., Garfinkel, S.L., Dahl, S.A., Dajani, A.N., Graham, M., Hawes, M.B., Karwa, V., Kifer, D., Kim, H., Leclerc, P., Machanavajjhala, A., Reiter, J.P., Rodriguez, R., Schmutte, I.M., Sexton, W.N., Singer, P.E., Vilhuber, L., 2020. The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau. Working Paper CED-WP-2020-009, U.S. Census Bureau, [Online]. Available: <https://www.census.gov/library/working-papers/2020/adrm/CED-WP-2020-009.html>.
- Alvim, M.S., van de Graaf, J., Gonze, R.G., Lemes, I.W., Nunes, G.H., 2020. TED 8750 Produto 02: Relatório Sobre os Riscos À Privacidade Decorrentes Da Atual Forma De Divulgação Dos Microdados Dos Censos Educacionais do INEP. Technical Report, Universidade Federal de Minas Gerais (UFMG) and Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), [Online]. Available: https://download.inep.gov.br/microdados/TED_8750-UFMG.pdf, In Portuguese. TED 8750 - PRICE Project. Report on privacy risks from the current disclosure model of INEP's Educational Census microdata.
- Alvim, M.S., Fernandes, N., McIver, A., Morgan, C., Nunes, G.H., 2022. Flexible and scalable privacy assessment for very large datasets, with an application to official governmental microdata. *Proc. Priv. Enhancing Technol.* 2022 (4), 378–399, [Online]. Available: <https://doi.org/10.56553/popets-2022-0114>.
- Andrade, D.F.d., Tavares, H.R., Valle, R.d.C., 2000. Teoria da Resposta ao Item: Conceitos e Aplicações. Associação Brasileira de Estatística (ABE), São Paulo, [Online]. Available: https://docs.ufpr.br/~aanjos/CE095/LivroTRI_DALTON.pdf, In Portuguese. Item Response Theory: Concepts and Applications.
- Baker, F.B., 2001. The Basics of Item Response Theory, second ed. ERIC Clearinghouse on Assessment and Evaluation, College Park, MD, [Online]. Available: <https://files.eric.ed.gov/fulltext/ED458219.pdf>.
- Barbaro, M., Zeller, Jr., T., 2006. A face is exposed for AOL searcher no. 4417749. *N. Y. Times* [Online]. Available: <https://www.nytimes.com/2006/08/09/technology/09aol.html>, Journalistic account documenting the re-identification of an AOL search-log user from pseudonymous query records.
- Brazil, 2011. Lei n. 12.527, de 18 de Novembro de 2011. Diário Oficial da União, Brasília, DF, [Online]. Available: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/112527.htm, In Portuguese. Brazilian Access to Information Law (LAI). Regulates access to public information.
- Brazil, 2018. Lei n. 13.709, de 14 de agosto de 2018. [Online]. Available: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/113709.htm, In Portuguese. Brazilian General Data Protection Law (LGPD). Presidency of the Republic.
- Brazil, 2024. Lei n. 15.017, de 12 de Novembro de 2024. Diário Oficial da União, [Online]. Available: <https://www2.camara.leg.br/legin/fed/lei/2024/lei-15017-12-novembro-2024-796541-publicacaooriginal-173508-pl.html>, In Portuguese. Amends Law No. 9, 394 (LDB) to mandate the publicity of educational data and microdata.
- College Board, 2024. Digital SAT Suite of Assessments Technical Manual. Technical Manual, College Board, [Online]. Available: <https://research.collegeboard.org/media/pdf/Digital%20SAT%20Suite%20of%20Assessments%20Technical%20Manual-FINAL.pdf>.
- Conselho Nacional de Saúde, 2016. Resolução no 510, de 07 de Abril de 2016. Diário Oficial da União, Seção 1, Brasília, DF, [Online]. Available: <https://www.gov.br/conselho-nacional-de-saude/pt-br/atos-normativos/resolucoes/2016/resolucao-no-510.pdf>, In Portuguese. Establishes ethical guidelines for research in Human and Social Sciences in Brazil.
- Daries, J.P., Reich, J., Waldo, J., Young, E.M., Whittinghill, J., Ho, A.D., Seaton, D.T., Chuang, I., 2014. Privacy, anonymity, and big data in the social sciences. *Commun. ACM* 57 (9), 56–63, [Online]. Available: <https://doi.org/10.1145/2643132>.
- De Capitani di Vimercati, S., Foresti, S., Livraga, G., Samarati, P., 2023. K-anonymity: From theory to applications. *Trans. Data Priv.* 16 (1), 25–49, [Online]. Available: <https://www.tdp.cat/issues21/vol16n01.php>.
- Franco, J.J., Miranda, F.L.d.A., Brancher, J.D., dos Santos, L.K., 2024. The main factors of ENEM: A literature and microdata perspective. In: 2024 IEEE Frontiers in Education Conference. FIE, IEEE, Washington, DC, USA, pp. 1–9, [Online]. Available: <https://doi.org/10.1109/FIE61694.2024.10893205>.
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2021. Exame Nacional do Ensino Médio (ENEM): Procedimentos de Análise. INEP, Brasília, DF, [Online]. Available: https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/enem_procedimentos_de_analise.pdf, In Portuguese. Brazilian National High School Exam (ENEM): Analysis Procedures.
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022a. Relatório de Impacto à Proteção de Dados Pessoais dos Microdados do Exame Nacional do Ensino Médio (ENEM). Data Protection Impact Report, INEP, Brasília, [Online]. Available: https://download.inep.gov.br/microdados/RIPD_dos_microdados_do_Enem.pdf, In Portuguese. ENEM Microdata Data Protection Impact Assessment (RIPD).
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022b. Nota de esclarecimento: Divulgação dos microdados. [Online]. Available: <https://web.archive.org/web/20260207173801/https://www.gov.br/inep/pt-br/centrais-de-conteudo/noticias/institucional/nota-de-esclarecimento-divulgacao-dos-microdados/>. (Updated 17 December 2025) Archived copy in Portuguese. Clarification note on microdata disclosure.
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2024. Microdados do ENEM 2023: Leia-me. INEP, Brasília, [Online]. Available: <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/enem>, In Portuguese. ENEM 2023 Microdata: Readme. Distributed within the official ENEM 2023 microdata package.
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2024. Microdados do ENEM 2024: Leia-me, In Portuguese. ENEM 2024 Microdata: Readme. Distributed within the official ENEM 2024 microdata package. [Online]. Available: <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/enem>.
- Justiça Federal da 1ª Região, 2025. Sentença tipo a. Processo no. 1027450-46.2022.4.01.3400. 8ª Vara Federal Cível da Seção Judiciária do Distrito Federal. Ação Civil Pública. PJe document ID 2215338599, In Portuguese. Merits judgment (Decision A).
- Justiça Federal da 1ª Região, 2026. Sentença tipo b (sentença integrativa). Processo no. 1027450-46.2022.4.01.3400. 8ª Vara Federal Cível da Seção Judiciária do Distrito Federal. PJe document ID 2228464760, In Portuguese. Judgment on the motions for clarification (Decision B).
- Ministério da Educação, 2012. Portaria Normativa no 21, de 5 de Novembro de 2012. Diário Oficial da União, Seção 1, Brasília, DF, [Online]. Available: <https://sisugestao.mec.gov.br/docs/portaria-2012-21.pdf>, In Portuguese. Regulation governing the Unified Selection System (SISU).
- Ministério Público Federal, 2022. Ação civil pública no 1027450-46.2022.4.01.3400. [Online]. Available: <https://web.archive.org/web/20240610001919/https://www.mpf.mp.br/df/sala-de-imprensa/docs/PRDFMANIFESTACAO147412022.pdf>, Procuradoria da República no Distrito Federal. 8th Federal Civil Court of SJDF, In Portuguese. Archived Public Civil Action filed by the Federal Public Ministry against INEP regarding microdata disclosure.
- Narayanan, A., Shmatikov, V., 2008. Robust De-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy. SP'08, IEEE, Oakland, CA, USA, pp. 111–125, [Online]. Available: <https://doi.org/10.1109/SP.2008.33>.
- National Center for Education Statistics, 2024. NAEP technical documentation: NAEP analysis and scaling. [Online]. Available: <https://nces.ed.gov/nationsreportcard/twd/analysis/default.aspx>.
- OECD, 2024. PISA 2022 Technical Report. OECD Publishing, Paris, [Online]. Available: <https://doi.org/10.1787/01820d6d-en>.
- de S. Paulo, F., 2018. Estudo inédito indica alta chance de fraude em mil provas do ENEM. [Online]. Available: <https://www1.folha.uol.com.br/educacao/2018/04/estudo-inedito-indica-alta-chance-de-fraude-em-mil-provas-do-enem.shtml>, In Portuguese. Journalistic investigation that analyzed anomalous ENEM answer patterns and cross-referenced the microdata with information on university entrants.
- Secretaria de Comunicação Social, 2024. Conheça o número de vagas do SISU 2024 por instituição. [Online]. Available: <https://web.archive.org/web/20260313125201/https://www.gov.br/secom/pt-br/acompanhe-a-secom/noticias/2024/01/conheca-o-numero-de-vagas-do-sisu-2024-por-instituicao>, In Portuguese. Archived official federal communication reporting 127 participating higher-education institutions.

- Shannon, C.E., 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423, 623–656.
- da Silveira, F.L., Barbosa, M.C.B., da Silva, R., 2015. Exame nacional do ensino médio (ENEM): Uma análise crítica. *Rev. Bras. Ensino Fis.* 37 (1), 1101, [Online]. Available: <https://doi.org/10.1590/S1806-11173710001>, In Portuguese. Critical analysis of the Brazilian National High School Exam (ENEM).
- Sweeney, L., 2000. Simple Demographics Often Identify People Uniquely. *Data Privacy Working Paper*, Carnegie Mellon University, [Online]. Available: <https://dataprivacylab.org/projects/identifiability/paper1.pdf>.
- Sweeney, L., 2002. K-anonymity: A model for protecting privacy. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems* 10 (05), 557–570, [Online]. Available: <https://doi.org/10.1142/S0218488502001648>.
- Wendler, C., Bridgeman, B. (Eds.), 2014. The Research Foundation for the GRE revised General Test: A Compendium of Studies. Educational Testing Service, [Online]. Available: <https://www.ets.org/pdfs/gre/gre-compendium.pdf>.
- Willrich, G., 2022. Plataforma para Visualização dos Microdados do ENEM (Bachelor's Thesis in Computer Science). Universidade Feevale, Novo Hamburgo, RS, Brazil, [Online]. Available: <https://biblioteca.feevale.br/Vinculo2/000042/000042c7.pdf>, In Portuguese. English title: “Platform for ENEM Microdata Visualization”.

All links were accessed on July 15, 2026.



Henrique Lindemann is an undergraduate student in Computer Engineering at the Federal University of Rio Grande do Sul (UFRGS), Brazil, and a Research Assistant at the Department of Electrical and Computer Engineering, RPTU Kaiserslautern-Landau, Germany. His research interests include data privacy, information security, and applied artificial intelligence.



Eder John Scheid is an Assistant Professor with the Department of Applied Informatics at the Federal University of Rio Grande do Sul (UFRGS), Brazil. He received his Ph.D. in Computer Science from the University of Zurich (UZH), Switzerland, in 2022. His research interests include blockchain systems, distributed systems, cybersecurity, cloud-native infrastructures, network and service management, and Network Function Virtualization (NFV).



Lisandro Zambenedetti Granville received the Ph.D. and M.Sc. degrees in Computer Science from the Federal University of Rio Grande do Sul (UFRGS), Brazil, in 2001 and 1998, respectively. He is a Full Professor at the Institute of Informatics of UFRGS and Director-General of RNP. His research focuses on network and service management, computer networks, and digital infrastructures. He has served in several leadership positions in the Brazilian computing and Internet communities.



Muriel Figueredo Franco is an Assistant Professor of Computer Science at the Federal University of Health Sciences of Porto Alegre (UFCSPA), Brazil. He received the Ph.D. degree in Computer Science (Summa cum laude) from the University of Zurich (UZH), Switzerland, in 2023, and was a Postdoctoral Researcher at the Federal University of Rio Grande do Sul (UFRGS), Brazil. His research interests include cybersecurity, digital health, communication systems, and applied artificial intelligence.