



Mapping Longitudinal Datasets on Educational Inequalities in Primary and Secondary Education: Research Findings and Implications

WP 3 - Deliverable 3.1 Mapping Research Report

Lead authors: Emmanuele Pavolini, Gianluca Argentin, Veronica Mobilio

Authors: Ilaria Lievore, Ekaterina Kochergina, Giulietta Zanga

Main Reviewers: Martin Neugebauer, Gabriele Ballarino

2025



Funded by the European Union's Horizon Europe Framework Programme (HORIZONCL2-2023-TRANSFORMATIONS-01-06) under grant agreement No 101132455. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



Funded by
the European Union



All LINEup resources are licensed under [CC BY 4.0 LEGAL CODE Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/)

Mapping Report Author Statement

Lead authors: Emmanuele Pavolini¹, Gianluca Argentin², Veronica Mobilio³

Authors: Ilaria Lievore³, Ekaterina Kochergina², Giuletta Zanga¹

Lead reviewers: Martin Neugebauer⁴, Gabriele Ballarino¹

Reviewer: Johanna Unewisse⁴, Luis Ortiz⁵, Giampiero Passaretta⁵, Kars van Oosterhout⁵, Léonard Moulin⁶, Priya Maurya⁶, Ioannis Cholezas⁷, Nikolaos Kanellopoulos⁷, Maria Fragkiadaki Theodorouleas⁸, Vincenzo Mauro¹⁰, Daniel Santos Da Costa¹¹.

Participating Institutions:

1. Università degli Studi di Milano La Statale (UNIMI - WP3 leader), Italy
2. Università degli Studi Di Milano-Bicocca (UniMiBicocca – WP3 leader), Italy
3. Fondazione per la Scuola della Compagnia di San Paolo (FpS – LINEup project coordinator), Italy
4. Karlsruhe University of Education (PHKA), Germany
5. Universidad Pompeu Fabra (UPF), Spain
6. Institut National D'Etudes Demographiques (INED), France
7. Centre of Planning and Economic Research (KEPE), Greece
8. University of Pireus (Pireus - WP2 leader), Greece
9. Computer Technology Institute and Press "Diophantus"
10. Università Degli Studi di Macerata (UNIMC), Italy
11. Universidade de Trás-os-Montes e Alto Douro (UTAD), Portugal

CRedit Authorship Contribution Statement¹:

Emmanuele Pavolini: Conceptualisation, Methodology, Validation, Formal Analysis, Writing - Original Draft, Writing - Review & Editing, Visualisation, Supervision, Project administration

Gianluca Argentin: Conceptualisation, Methodology, Validation, Formal Analysis, Writing - Original Draft, Writing - Review & Editing, Visualisation, Project administration, Supervision

Veronica Mobilio: Conceptualisation, Methodology, Validation, Formal Analysis, Writing - Original Draft, Writing - Review & Editing, Visualisation, Supervision, Project administration, Funding acquisition

Ilaria Lievore: Methodology, Validation, Investigation, Data Curation, Writing - Original Draft, Visualisation

Ekaterina Kochergina: Validation, Investigation, Data Curation, Writing - Original Draft, Visualisation

Giuletta Zanga: Validation, Investigation, Data Curation, Writing - Original Draft

Infographic Layout: Andrea Succhielli, European Grants International Academy (EGInA)

Corresponding author: Veronica Mobilio, Piazza Bernini, 5 10138 Torino - Italy, +39 011 1563 0500, veronica.mobilio@fondazione scuola.it

Acknowledgement of Contributors to Dataset Mapping: the authors gratefully acknowledge the researchers and institutional teams who contributed to the identification, collection, and documentation of the longitudinal datasets analysed in this report. Contributors include researchers from the eleven institutions listed above.

How to cite:

Pavolini E., Argentin G., Mobilio V., Lievore I., Kochergina E., Zanga G. (2025). *Mapping Longitudinal Datasets on Educational Inequalities in Primary and Secondary Education: Research Findings and Implications*. Deliverable D3.1, Horizon Europe Project LINEup.

¹ CRediT Authorship: <https://www.elsevier.com/researcher/author/policies-and-guidelines/credit-author-statement>

Executive summary

Introduction

Longitudinal data offer valuable insights into students' learning trajectories and how educational inequalities emerge, persist, or change over time. While many European countries have invested in education data collection, longitudinal datasets specifically tracking student learning outcomes over time remain fragmented and inconsistently documented. The LINEup project addresses this gap by conducting the first comprehensive mapping of existing national and regional longitudinal datasets on student learning outcomes at the primary and secondary education levels (International Standard Classification of Education - ISCED 1, 2, 3).

The mapping covers 32 countries: the 27 EU Member States, three European Economic Area (EEA) associated countries (Iceland, Liechtenstein, and Norway), as well as Switzerland and the United Kingdom. This selection reflects the primary geographic scope of the LINEup project, which focuses on European education systems while ensuring representation of diverse policy frameworks, administrative structures, and data collection traditions within the broader European context. The inclusion of EEA countries, Switzerland, and the UK allows for a more comprehensive understanding of longitudinal data availability and use beyond the EU alone. In addition to this core group, selected practices from outside Europe - particularly from the United States, displaying leading examples in the field - were considered and reviewed during the methodological design to enrich the project's comparative lens.

This report presents the results of the mapping exercise, conducted under WP3 of the LINEup project. The mapping aimed to identify, document, and classify existing datasets; evaluate their analytical relevance for studying educational inequalities; and explore their potential for cross-country comparability. The mapping also provides the empirical foundation for two key project outputs: the *LINEup Education Data Explorer*, an online platform that supports dataset discovery and reuse by researchers and educational stakeholders; and a *Feasibility Study* on data harmonisation (Deliverable D3.2 due in January 2026), which will assess the conditions for aligning and comparing data structures and indicators across countries. Together, these outputs aim to foster a stronger and more integrated use of longitudinal data in education research and equity-oriented policy development.

Methodology

The mapping followed a structured, multi-step strategy grounded in a shared theoretical and methodological framework. It built upon three main sources:

- A **systematic literature review** (conducted in WP2), which identified 43 relevant datasets used in empirical studies on educational inequalities with a longitudinal or repeated cross-sectional research design (out of the 77 collected overall).
- An **expert questionnaire** sent to over 300 stakeholders across 28 countries, which led to 70 additional dataset suggestions.
- **Structured web searches and desk research** conducted by project partners across 32 countries, complemented by collaboration with sister projects, which added a further 116 datasets to the inventory.

All suggested datasets were reviewed using predefined inclusion criteria: they had to be longitudinal; refer to primary (ISCED 1) and/or secondary education (ISCED 2, 3); focus on student learning

outcomes (such as standardized achievement tests or school grades); originate from one of the 27 EU Member States, EEA associated countries, Switzerland, or the UK; be nationally or regionally representative (or based on a large scale sample); and be ongoing or recently collected (with the latest wave not older than 2014).

Eligible datasets were documented using a standardised online form that captured key metadata on basic dataset information (name, responsible institution, country, accessibility conditions) along with details on longitudinal structure, grade coverage, data type, skills and educational outcomes, background variables related to student, household, teacher and school characteristics.

A structured validation protocol was applied to ensure relevance, accuracy, and comparability of the documented datasets. This included an initial quality review by the mapping team, followed by peer validation by at least one additional project partner. Datasets flagged as *uncertain* or *borderline* were subject to manual re-assessment by the WP3 coordination team. Finally, all validated datasets were reviewed a second time to ensure cross-country consistency and alignment with inclusion criteria and classification guidelines. The entire validation process was grounded in a transparent, replicable methodological framework, ensuring coherence across countries and robustness in comparative dataset classification.

Key Findings

Of the 229 datasets identified, 104 met the inclusion criteria and were included in the final mapping. These datasets are distributed across 24 of the 32 countries covered by LINEup. The countries with the highest number of eligible datasets are Germany (31) and the United Kingdom (11), followed by Finland and Portugal (both 7), and France (5). A clear geographic pattern emerges: most countries in Western and Northern Europe have at least one eligible dataset, often stemming from well-established administrative systems or long-standing student assessments. In contrast, several countries in Central and Eastern Europe have no eligible datasets, reflecting more limited investments in longitudinal data infrastructure and/or lower levels of data accessibility. These patterns often mirror national policy priorities, institutional capacity, and the overall maturity of student assessment or data collection systems.

The mapped datasets reveal substantial variation in design, scope, coverage, and accessibility. Nearly half (46.6%) rely exclusively on survey data, while a smaller share (19.4%) is based solely on administrative sources. An additional 34% adopt a hybrid approach that combines both survey and administrative information. Although administrative datasets often reflect national assessment systems and official registries, typically covering larger share of the population and longer timespan, survey-based and hybrid datasets tend to provide richer contextual and background information, particularly on student characteristics, household conditions, and school environments. This is largely due to their research-oriented design.

The longitudinal span of the datasets ranges from 2 to over 14 years, reflecting diverse research objectives and data collection strategies. While many datasets follow students for five to eight years, longer-term tracking remains relatively rare. Nearly half of datasets (47.1%) cover both primary and secondary education, while the remaining focus exclusively on either primary (21.2%) or secondary education levels (31.7%). Coverage is often concentrated around key transition points - particularly grades 8, 9, and 10 - rather than providing continuous observation across all educational stages. Notably, around 30% of datasets begin in pre-primary education, enabling early tracking of inequalities before formal schooling. While several datasets include data points in upper secondary education (grades 11–13), continuous individual-level tracking across the full span of compulsory

schooling is relatively rare. The few datasets that provide this level of coverage, typically designed as birth cohorts, remain the exception.

In terms of outcomes, the vast majority of datasets include academic performance measures in the form of standardised test scores and/or teacher-assigned grades in literacy and numeracy. Fewer datasets (around 30%) cover either science, foreign languages, or other skills domains. School- and teacher-level variables are inconsistently reported: only a minority of datasets include classroom-level information or teacher-level data.

Data linkability is uneven across the mapped datasets. Only 29.1% are both linkable and accessible to external users; an additional 26.2% allow linkage but restrict access to data owners, while 44.7% does not allow linkability to other data sources at all. True inter-dataset linkage - such as connecting learning outcomes to enrolment records or external registries - is generally limited to countries with strong administrative systems, such as Germany or the Nordic countries. Legal, technical, and institutional barriers continue to hinder broader data integration and reuse, particularly where student identifiers are absent, or governance frameworks are fragmented.

Finally, access to microdata represents a major constraint. Only 10.7% of datasets are publicly accessible (open data), while 49.5 % require an ad hoc request and formal authorisation to obtain full access. A further 26.2% are either inaccessible or subject to significant restrictions. In addition, metadata on accessibility conditions are unclear or not available for a consistent share of datasets (13.6%). Overall, metadata have been often reported to be incomplete, outdated, or inconsistently documented - limiting transparency and reducing the potential for secondary analysis and evidence-based policy use. The combination of limited accessibility, documentation gaps, and divergent metadata standards reduce the usability of longitudinal education data across the mapped countries. Moreover, while the diversity of aims, structures, and target populations across datasets reflects the richness of the field, it also entails potential barriers that researchers and policymakers may face when trying to identify the most suitable sources for specific analytical or strategic purposes.

To support dataset analysis, harmonisation, and comparative use, a structured classification framework was developed. It organises the mapped datasets according to data structure, outcome types, availability of equity-related variables, and longitudinal tracking potential. This framework also underpins the visualisation and filtering functionalities of the *LINEup Education Data Explorer*.

Discussion and Conclusions

The mapping reveals both opportunities and limitations in the current landscape of longitudinal education data in Europe. On the one hand, the 104 datasets included in the mapping provide a unique, albeit uneven, empirical foundation for analysing student learning trajectories and educational inequalities. On the other hand, persistent disparities in data access, scope, and usability hinder broader analytical and comparative potential.

Several key insights emerge:

- **Data exist but are underused and frequently inaccessible:** Although numerous longitudinal datasets have been developed across Europe, they remain underutilised due to a combination of structural and procedural constraints. In several cases, even high-quality administrative data are off-limits to researchers because of the absence of clear data-sharing protocols, insufficient institutional coordination, or overly cautious interpretations of data protection regulations.
- **Significant heterogeneity in variables collected:** While most datasets focus on academic outcomes, particularly literacy and numeracy, the coverage of other dimensions - such as

non-cognitive skills, school climate, teaching practices, or equity-related indicators - varies widely. Only few datasets collect repeated measures on socio-emotional outcomes or classroom-level dynamics, which are essential for understanding broader learning conditions.

- **Limited comparability across countries:** Despite the presence of valuable data infrastructures in many countries, the lack of harmonised definitions, coding schemes, and measurement instruments limits cross-country comparison. For example, similar constructs (e.g., parental education) are measured differently across datasets - even within the same country - complicating efforts to build shared indicators or conduct joint analyses.
- **Uneven quality and granularity:** The granularity of information also varies. While some datasets enable longitudinal tracking at the individual level across multiple transitions, others are restricted to a single education stage or provide only aggregate data. This limits the scope of research questions that can be addressed, particularly those related to equity trajectories, long-term outcomes, or causal inference.
- **Technical and legal barriers to linkage:** Although linkability across datasets exists in several contexts - especially where strong administrative systems are in place - it remains underexploited. Legal, technical, and institutional barriers hinder more systematic integration, particularly in countries where centralised student identifiers are absent and governance frameworks for data management remain fragmented or underdeveloped.

Despite these limitations, the mapping lays the groundwork for improved use and potential harmonisation of education data in Europe. It enables researchers and educational stakeholders to identify promising datasets for future analysis and to draw inspiration from exemplary practices. While other countries - such as the UK or Finland - also host well-developed longitudinal datasets, the Italian INVALSI-SNV system and the German NEPS programme were selected as illustrative cases due to their comprehensive design, longitudinal depth, inclusion of contextual variables, and accessibility protocols. These features closely align with the core dimensions of the LINEup mapping framework, illustrating how coordinated investment in data infrastructure and governance can foster robust and policy-relevant research on educational inequalities. Beyond Europe, the U.S. Early Childhood Longitudinal Study (ECLS) provides an additional benchmark: by tracking the same individuals over time and systematically linking student data with information on families, teachers, and schools, it exemplifies the potential of fully integrated longitudinal systems to support developmental research and policy evaluation. The existence and use of rich longitudinal datasets in several countries can serve as a catalyst for increased awareness and commitment among national authorities in less advanced contexts, encouraging progress in data infrastructure and governance.

Implications for Research, Policy, and Practice

The results of the mapping exercise have important implications across multiple domains.

For researchers: The validated inventory of 104 datasets, supported by a structured classification framework and rich metadata, offers a valuable starting point for future comparative studies and secondary analyses and for informing the design of future data collections that leverage the strengths of existing datasets. By enabling dataset filtering by type, coverage, and variables, the mapping facilitates more informed research design and supports future harmonisation efforts. It also highlights best practices that can inform methodological innovation and cross-country collaboration.

For policymakers (including national and local authorities): The mapping enhances transparency regarding the availability and quality of longitudinal education data across Europe. It highlights significant gaps in data infrastructure - especially in Central and Eastern Europe - and identifies barriers to data accessibility, including restrictive authorisation regimes or insufficient documentation. These insights are critical for designing policies that are both evidence-based and responsive to equity challenges. The report also emphasises the need for coordinated investment in data systems as a strategic asset for education governance. Finally, it draws attention to severe limitations in the capacity to link existing datasets with administrative sources - an obstacle that hampers efforts to understand the complex mechanisms underpinning the development of educational inequalities.

For education systems and practitioners: The mapping facilitates the identification of key sources of empirical evidence on education inequalities, supporting the design of evidence-based interventions. Its findings highlight the importance of expanding data collection beyond academic outcomes to include contextual, school-level, and non-cognitive dimensions (e.g., school engagement, motivation, well-being). Without such information, it is difficult to monitor progress toward more inclusive and equitable learning environments, or to evaluate the impact of policies and practices aimed at reducing educational inequalities. The mapping also underscores the value of linking datasets with different scopes and information, within and across systems, to support a more holistic understanding of students' educational experiences and outcomes.

To design effective and equitable education policies, policymakers need access to timely, high-quality longitudinal data. Promoting a culture of data use and addressing the structural, legal, and technical barriers to access, interoperability, and documentation is essential to improve the effectiveness, equity, and accountability of education systems across Europe. Doing so will empower researchers, decision-makers, and educational stakeholders to generate actionable insights based on robust data, and co-design strategies to reduce educational inequalities.

Report Structure

Section 1 introduces the rationale, objectives, and research questions behind the mapping. Section 2 outlines the shared theoretical and methodological framework, detailing data sources, inclusion criteria, validation procedures, and the dataset classification system. Section 3 presents the results of the mapping, including the number, type, and distribution of eligible datasets. Section 4 provides a descriptive analysis of the mapped datasets, highlighting their research potential, structure, and cross-country comparability. Section 5 introduces the *LINEup Education Data Explorer*, describing its design and key functionalities for dataset discovery and use. Section 6 discusses the feasibility of data harmonisation, identifies common challenges, and lays the groundwork for the upcoming *Deliverable D3.2*. Section 7 concludes the report with key findings and recommendations. After the report bibliography (section 8), the Annexes include, in order: the full list of included and excluded datasets, the expert questionnaire, the online form for dataset metadata collection, and a glossary of technical terms.

Infographic - Visual Summary of Key Findings

The infographic below provides a visual synthesis of the key insights from the mapping exercise. It summarises the number and types of datasets identified, their scope, structure, and content, as well as main barriers to access and use. It also illustrates what existing longitudinal data in Europe can and cannot support in the study of educational inequalities.

PROJECT OVERVIEW

The LINEup project conducted
**the first EU-wide mapping
of existing longitudinal datasets**
on student learning outcomes
within primary and secondary school education.

104 Datasets Mapped*



24 European Countries (out of 32)



Focus: Student Learning Outcomes
and Academic Performance

Data Types: Longitudinal
(i.e. panel, repeated cross-sections,
pseudo-panels)

*One U.S. dataset is included as a best-practice example,
but excluded from the descriptive statistics

RESEARCH POSSIBILITIES

The potential of a dataset depends on the research questions
we ask. Most support descriptive research on individual traits.
But explaining inequalities through school-level
or comparative analysis remains difficult. Even when access
is granted, poor documentation limits reusability.

✓ WHAT WE CAN DO

Research Possibilities

- Analyse learning trajectories
(especially via test scores
or grades)
- Study socio-demographic
effects
(presence of basic ascriptive
variables)
- Track grade retention
and transitions
(coverage of key educational
transitions)
- Compare performance
by school type/track
(availability of basic
school-level identifiers)

✗ WHAT WE CANNOT DO

Analytical Limitations

- Study peer/class composition
(lack of class-level data)
- Assess impact of teacher
quality or turnover
(limited variables)
- Analyse school segregation
or territorial inequalities
(no geo-referencing)
- Compare non-cognitive
skills cross-country
(heterogeneous or missing
data)

METHODOLOGY | A Three-Step Strategy

1. DATASET IDENTIFICATION

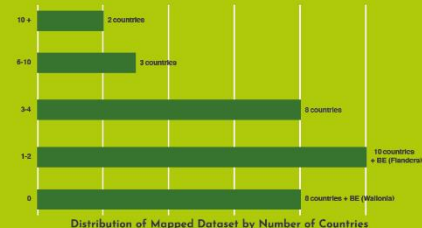
- Literature Review (**43**)
- Expert Questionnaire
sent to 315 Experts
from 28 Countries (**70**)
- Targeted Web-Search
& Partner Knowledge (**108**)
- Collaboration with
other projects (**8**)

2. DATASET SELECTION

- 229** datasets **screened**
- 104** datasets **included**
- 125** datasets **excluded**

3. DATASET ANALYSIS

Focus on Data Structure, Time
Span, Indicators & Variables,
Quality, Accessibility



KEY FINDINGS | By the numbers

HOW are datasets designed, linkable, and accessible to users?

- 68%** are panel data
(individual tracking over time)
- 22%** are repeated cross-sectional
(with individual identifiers)
- 10%** can be used as pseudo-panel
- 47%** are survey data
- 19%** are administrative data
- 34%** are hybrid data (survey + admin)
- 55%** allow linkage to other data sources
- 11%** are open access
- 50%** are accessible upon request
- 26%** are highly restricted or not accessible

WHO is included in the datasets?

- 32%** cover the entire student population
- 45%** use random samples
- 23%** use non-random samples
- 50%** with sample < 5,000 students
- 29%** with sample > 10,000

WHAT do the datasets measure?

- 77%** include standardised tests
- 69%** include school grades
- 53%** include both tests and grades
- 64%** cover both numeracy
and literacy skills
- 36%** cover foreign language skills
- 30%** cover science skills
- 69%** cover other domains skills
- 95%** include information
on students' gender
- 68%** include information
on foreign birth country
- 74%** include information
on parental education
- 61%** include information
on parental occupation
- 30%** include information on teachers
- 23%** include information
on classroom composition

WHERE do the datasets apply?

- 62%** have national coverage
 - 38%** are sub-national.
- Most datasets come from Western
& Northern Europe

WHEN were data collected and for how long?

- 23%** span 14+ years
 - 20%** cover only 2 years
 - 51%** are still ongoing
 - 47%** cover both primary and secondary
education.
- Most frequently: grades 8, 9, 10
30% begin in kindergarten

The value of longitudinal
data lies in its responsible,
effective use.
LINEup contributes
to a more transparent
and visible education
data landscape in Europe.

