Coherence of integrated data from the Italian Education and Training Register in the framework of the Statistical System of Registers

Giovanna Brancato, Claudia Busetti, Lucia Coppola

Istat – Italian National Statistical Institute, Italy

Abstract

Istat is developing a new statistical Thematic Register on Education and Training (TRET), based on administrative data, properly arranged to provide official yearly statistics, as well as longitudinal information on individual educational patterns. The first release is planned in 2026. Being part of the Istat Integrated System of Statistical Registers, the TRET has to guarantee consistency with the registers on the base populations and their characteristics, namely the Base Statistical Register of Individuals and the Base Statistical Register of Economic Units. The former provides time invariant (gender at birth, birth date and place) and variant (citizenship, residency) variables, while the latter provides the most relevant variables on education and training institutions. The process of building up the TRET involves a set of operations mainly aimed at improving the quality of the input data. Firstly, education and training variables from input data are validated. Inconsistency of individual characteristics are checked and edited. Secondly, the integration with the base registers allows assigning validated demographic variables to the individuals but requires the treatment of data inconsistency between the base register variables and the administrative input data. Finally, TRET estimates are computed from the core statistical unit *education position*, identified by the triplet of elementary units: the *individual*, the *education and training institution* and the *education and training program* in which the student is enrolled. Further editing and imputation is adopted to ensure cross-sectional and longitudinal consistency of these derived units. In this paper, the process of implementation of the annual TRET is described with reference to the grades up to the upper secondary education. All the typologies of process quality controls are illustrated and examples of estimates of quality indicators are provided. Finally, indicators on coherence of the estimates with other sources are reported.

Keywords: data consistency, coherence of estimates, integrated data, quality measures

1. Introduction

A new statistical Thematic Register on Education and Training, named TRET, is under development at Istat (Brancato *et al.*, 2023). The register is fed with all the available administrative data on education and training, properly re-organised for statistical purposes. Most of the administrative data are supplied by the Ministry of Education and Merit (MIM) and the Ministry of University and Research (MUR). In addition, the National Institute of Documentation, Innovation and Education Research (Indire) provides data on tertiary short cycle programs, i.e. data on the Superior Technological Institutes Academy (ITS Academy). Finally, the National Institute for the Evaluation of the education and training system (Invalsi) provides data on standardised performance tests in the national scholastic system.

The TRET has an annual and a longitudinal component. The time reference for the estimates is the scholastic/academic year for the enrolments and the calendar year for licensed/graduated students. In short, the main register-based statistical estimates concern:

- schools and their characteristics (private, public, school level, equipment, …);
- other institutions (universities, foundations, …) providing education and training and their characteristics (e.g. the legal form);
- enrolled students in every grade, from pre-primary to university education by education and training program;
- dropouts and students repeating the grade;
- students attaining primary, lower and upper secondary licenses; bachelor, master and PhD graduates by education and training program;
- students changing education and training program within the same or between different scholastic/academic year(s);
- standardised (Invalsi tests) and unstandardised (state examinations, graduation and bachelor scores) performance;
- teaching staff and its characteristics.

Longitudinal data allow analysing individual education and training paths and outcomes, taking into account for contextual factors (e.g. variables on the socio-economic conditions).

Data of the register are organised into different and linked information areas. The core of the register is the area on the statistical unit called *education position*, i.e. a derived unit formed by three elementary units: the *individual*, the *institution* (school/university) and the *education and training program*. The derived unit allows tracking each individual education spell, including changes of school/university and/or program. TRET manages education and training programs at the maximum detail available from the input sources, mapped with the International Isced and Isced-F classifications throughout the new national classification of the education and training programs and attainments (Brancato & Grassi, 2021). Other areas of the register concern: *i)* the unit *education and training institution* and its characteristics (public or private school, university or foundation); *ii)* the *qualifications* attained by each individual and the relevant available information (e.g. the gained scores); *iii)* the results of Invalsi tests carried out during the education history (grades 2, 5, 8, 10, 13); *iv) internships* done within ITS Academy programs; *v)* the *teaching staff* and its characteristics.

The TRET is part of the Istat System of Statistical Registers (ISSR) (Radini *et al.*, 2018) and is constrained to information on the units of the Base Register on Individuals (RBI) and the Base Register of the Economic Units (RBUE). The TRET is supported by a system of metadata and quality documentation compliant with the most recent Istat developments (Di Zio *et al.*, 2023). TRET estimates have to fulfil coherence criteria both within the ISSR and with respect to data published by the owners of the administrative sources.

In this paper, the annual TRET process is described in Section 2. Sections 3 and 4 present the results of the analyses on internal (within the ISSR) and external (with data published by the MIM) coherence, respectively, with reference to 2021-2022 scholastic year. Finally, some conclusions are drawn in Section 5.

2. Statistical process and process quality

In this section, the process and the quality controls developed to build the annual version of the TRET on school education (from pre-primary to upper secondary education) are described. Input data are organised into three datasets: a) enrolments (some students may have more enrolments); b) licensed/graduated students; c) schools. These datasets are integrated and linked with Istat *base registers* (RBI, RBUE). In compliance with Istat privacy rules, data on persons are treated in an anonymised form, and linkable using pseudonymous codes, centrally assigned to each unit. The sources a) and b) undergo the following Generic Statistical Business Process Model (GSBPM) sub-process steps, marked with the numbering they have in the published model (Unece, 2019).

5.2. Classify and code. The territorial variables (birthplace, school municipality) are checked and missing values are imputed. Scholastic variables are checked, however so far missing information have not been detected. The education and training programs (e.g. *Scientific Lyceum* or *Industrial Technology* program within the upper secondary education), are harmonised and linked with the national classification assigning a standard and unique code.

5.3. Review and validate. Four main steps are applied: deduplication, logical deletion of not usable records, within-record and within-cluster[1] consistency controls. First, duplicate records with respect to the variables not relevant for the building up of the register, as well as records with missing information on the variables used to derive the *education position unit*, are marked for deletion. Records are then checked respect to inconsistencies in the scholastic variables. Within-record controls (e.g. incompatible school level and grade) and within-cluster checks (e.g. signals of enrolment in different grades in the same year) are applied. Records in clusters with data inconsistencies are marked for deletion, constraining to the maintenance of at least one record per cluster.

[1] A cluster is a set of records concerning the same student. Since a student may change school and/or program during the year, some individuals can have more records.

5.4. Editing and imputation. Within-record and within-cluster inconsistencies are resolved, based on deterministic rules and with the support of data from the previous scholastic year.

5.1. Integrate data. Data from sources a) and b) are integrated using as key variables the anonymised individual code, the school and the education and training program code. Source c) is integrated by means of the school code. Deterministic linkage is adopted.

5.4. Editing and imputation. Following the integration, part of the incoherencies between enrolment and license/graduation data are resolved, while others are left for the following treatment step, which includes additional information supporting the decision.

5.5. Derive new units and variables. The statistical unit *education position* formed by the triplet of elementary units: *individual*, *school*, *education program* is created. If one of them changes, the current education position closes and a new one is opened. The variable assigned to the *education position unit* are: starting and ending dates, entry and exit reasons, enrolment status (yes or not) especially useful for clusters. Quality controls checks are applied to verify and resolve possible conflicting situations. Supporting metadata are built.

5.1. Integrate data. Integration with RBI using the anonymised individual code is carried out. Differences between the TRET population and the correspondent subset of RBI individuals are investigated (see Section 3). Integration with RBUE is also performed (not shown in this paper).

5.7. Calculate aggregates. Estimates of the number of enrolled and licensed/graduated students by gender, citizenship, school level, etc. are computed. Other estimates, such as the dropouts, can only be partially counted, since the correct methodology for measuring the dropouts requires the analysis over two scholastic years.

6.2. Validate outputs. Estimates from the TRET are compared to data published by the MIM (see Section 4). The results are preliminary since they are based on the annual process of the TRET, before the longitudinal version is created and data are longitudinally reconciled.

3. Coherence within the Istat System of Statistical Registers (ISSR)

The coherence within the ISSR is evaluated with respect to the base unit *individual*. In the ISSR, the RBI provides all the *thematic registers*, e.g. the TRET, with the demographic variables. However, the administrative data used for the TRET include demographic variables that may show differences from those available from RBI. Therefore, an assessment of the impact of the integration with the RBI is needed in order to be aware of:

- to what extent students in TRET are not represented in RBI and the differences between the variables available in both sources (Section 3.1);

- the amount of RBI individuals in compulsory school age not in the TRET (Section 3.2).

TRET 2021/2022 data are linked with RBI referred to December 2021. RBI provides information on resident individuals (almost 60 million of individuals at the end of 2021), and on non-resident individuals found in the administrative sources populating the register, leading to a database of about 110 million of records in total in the same year. We take into account the whole population provided by RBI to gather the demographic information of interest. For the sake of simplicity, we consider only the most relevant time-constant variables: sex, age[2] and place of birth.

3.1 Linkage of TRET students with RBI population

According to the TRET, the students' population in 2021/2022 is made of 8,366,912 individuals attending pre-primary, primary, lower and upper secondary school. Individuals with a single record or with a cluster of records are analysed separately.

Most of the students are represented in RBI, with the exception of 12,420 individuals (0.15%). Although this is a negligible percentage of students, it is worthy analysing their characteristics. These students are equally distributed among sex. In comparison with students linked with RBI, they are over-represented in the primary school age class (6-10 year old) and under-represented in the upper secondary school age class (14-18). Furthermore, most of them are born in a foreign country (77% *vs* the 4.6% of students found in RBI) and show incomplete information about the place of residence (17%).

Table 1: Agreement rates of demographic variables between TRET and RBI

Variable	Agreement Rate
Gender	99.81
Year of birth	99.99
Italian province and municipality birthplace	98.04
Foreign country birthplace	93.79

The comparison of the demographic variables in TRET and RBI show very large accordance. Table 1 shows the agreement rates (i.e. the percentage of individuals equally coded in both sources). Only the variables on the birthplace show a slightly lower level of consistency (98%

[2] We show the comparison considering only the year of birth, because differences in terms of day or month of birth do not affect official estimates on education.

on the Italian province and municipality and about 94% on the foreign country of birth). About 2% of records need an editing and about 0.1% an imputation of at least one of the selected demographic variables to let the TRET consistent with RBI (see Table 4 in Appendix 1).

3.2 RBI student population not found in TRET

Another explored issue concerns the RBI resident population in compulsory school age (6-16 year old) that is not represented in the TRET. About 300,000 children who do not have signals in the TRET were identified, representing around 5% of the students enrolled in the compulsory school according to administrative data (5,820,157). Half of them are aged 13 or over. Although most of them are born in Italy, there is a high incidence of children born abroad (20%). The incidence is lower for 6-13 year old students (i.e. 3.5% for children aged 6-10 and 3.2% for children aged 11-13), and higher (9.3%) for 14-16 year old students. This may partially be due to the existence of students enrolled in vocational regional programs (about 90 thousands individuals in this age class) who are not covered in the TRET, due to the unavailability of the microdata source. Disregarding these students, the incidence becomes similar to that in the other age classes (i.e. 3.6%) (see Table 5 in Appendix 2).

4. Coherence with external sources

An important quality issue concerns the coherence of the estimates that can be derived by the TRET respect to those published in MIM open data[3], which are computed from the same administrative source transmitted to Istat. MIM disseminates statistics on a subset of schools (some *non-state-recognised* private schools are excluded) and on the *attending students* (the dropouts and the students moving abroad or out of the *state-recognised* scholastic system during the year are excluded). In addition, in the pre-primary schools also some special *early classerooms* with children aged three or less are excluded by MIM. Consequently, data from TRET are selected to meet the MIM definitions. In addition, two autonomous Italian regions (Trentino Alto Adige, Valle D’Aosta) where excluded in both sources due to incompleteness in the data transmitted to Istat. These exclusions lead to a decrease of the total students’ population to 8,068,103. Table 2 reports the number of *schools* and *attending students* estimated from the TRET and the percentage relative variations[4] respect to MIM estimates. Comparisons by gender and citizenship, variables constrained to RBI information, are also presented. As shown, the differences in the estimates of the *schools* and *attending students*

[3] https://dati.istruzione.it/opendata/opendata/catalogo/#Scuola

[4] The relative variation is computed as the difference between the TRET and the MIM estimates, relative the MIM estimate per cent.

are close to zero in almost all schools' levels. The maximum detected difference is 1.3%. Regarding the number of schools, differences in the upper secondary level may be due to misclassified *state-recognised* school.

Table 2. Number (n) of schools and attending students estimated from the TRET by gender and citizenship*, and percent relative variations (r.v.) respect to MIM open data, for school level

	Pre-Primary		**Primary**		**Lower Secondary**		**Upper Secondary**	
Unit	**n**	**r.v.**	**n**	**r.v.**	**n**	**r.v.**	**n**	**r.v.**
Schools	21,249	0.0	16,010	0.0	7,835	0.0	8,002	0.1
Attending students	1,291,184	1.3	2,467,420	-0.1	1,647,672	-0.1	2,661,827	-0.1
Males	665,889	0.8	1,270,212	-0.1	851,271	-0.1	1,367,455	-0.2
Females	625,295	1.8	1,197,208	0.0	796,401	0.0	1,294,372	0.0
Italians	1,125,662	0.5	2,169,489	0.8	1,481,528	1.5	2,480,036	1.2
Not Italians	165,513	7.1	297,923	-6.3	166,140	-12.4	181,783	-15.1

* Trentino Alto Adige and Valle D'Aosta regions are not included. The gender and citizenship are taken from RBI when available, otherwise from the TRET. Some missing values remain for the citizenship.

Considering the number of *attending students*, the difference for the pre-primary education is attributable to TRET inability to subtract the children in *early classrooms* (information not available in the data transmitted to Istat), therefore is due to the adoption of different definitions. Indeed, if the age of the children is considered as proxy measure of attendance in *early classrooms*, the estimates become even closer. However, the age does not identify exactly children in *early classrooms* as the 3 year old children can attend also regular classes. Small differences are also confirmed with respect to the students by gender. As expected, the analysis for citizenship highlights larger differences that, however, have a limited impact.

5. Conclusions

In the framework of the ISSR, TRET data have to fulfil a requirement of internal coherence respect to the RBI population. Our analysis showed a negligible amount of the students in TRET not listed in RBI. They are characterised to be more likely younger and born abroad, leading to the hypothesis that this population could be missed in RBI. In fact, due to incompatibility with its production scheduling, the administrative data on the schools' students are not integrated in RBI, and this explains the estimated under-coverage. Consequently, in order to bridge this gap, RBI is going to adjust its production process. In addition, the results showed that there is a quota of resident individuals, in compulsory school age (6-16 year old) listed in RBI and not in TRET. Some hypotheses on this quota have been formulated. On the one side, it is known that the administrative sources used in TRET are affected by under-coverage (vocational regional students, students in foreign schools in Italy, some not *state-recognised* schools). On the other side, RBI may be affected by some over-coverage, as we verified that around 6% of the above mentioned quota become not resident in 2022. The rest includes individuals who do not enrol in the school system (dropouts before 2021). Considering the quality of the estimates, the results on coherence with MIM figures are encouraging. This is not a trivial conclusion, as data acquired by Istat undergo several treatments, aimed at improving the quality and managing the information for statistical purposes. It is worth mentioning that these analyses are based on the annual version of the register, before the longitudinal setting and adjustments. Thus, the results need to be further confirmed after the implementation of the longitudinal version of the TRET.

Acknowledgment

The authors thank all the Istat members of the Task Force in charge of the planning and implementation of the TRET and the colleagues from the MIM for the fruitful discussion and share of views on these preliminary results.

References

Brancato, G., Grassi, D., Busetti, C. (2023). Towards a system of integrated statistical data on education and training. *Proceedings of Statistics Canada Symposium 2022* https://www150.statcan.gc.ca/n1/en/pub/11-522-x/2022001/article/00009-eng.pdf?st=PTQU8PrB

Brancato, G., Grassi, D. (2021). Towards an integrated system for the production of relevant statistical data on education and training. *VI Seminar "INVALSI data: a tool for teaching and scientific research"*. Chapter 1. Franco Angeli ed. (*accepted for publication*)

Di Zio M., Falorsi S., Rocci R., Simeoni G. (2023). Process and output quality evaluation measures for Istat Integrated System of Statistical Registers EESW 23, Lisbon, 20-22 September 2023

Radini, R., Scannapieco, M., Tosco, L. (2018). The Italian Integrated System of Statistical Registers: Design and Implementation of an Ontology-based Data Integration Architecture. https://www.istat.it/it/files/2018/11/Scannapieco_original-paper.pdf

Unece (2019). Generic Statistical Business Process Model GSBPM v.5.1. https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

Appendix 1

The integration of TRET with RBI may: *i)* not to introduce any change in the demographic variable of interest when the sources agree (Table 1); *ii)* to correct the variables when inconsistent; *iii)* to impute the missing values in TRET. Before correcting or imputing a variable of interest, we have to consider the whole set of demographic variables taken from RBI, to assess if it is “fully correct” or not. We consider the whole set as fully correct if all the variables are complete and coherent (e.g. if at least one variable is missing, the set is not considered as fully correct; if the foreign country of birth is provided together with the Italian province and municipality of birth, it is not considered as fully correct). If the set is not fully correct, the integration with RBI is questionable because it may reduce the consistency of the demographic variables in the TRET.

According to Table 4, in most of the cases, the integration does not have any effect (about 98%). This evidence is relevant meaning that the administrative data used to populate TRET appear mostly reliable in terms of the demographic information. The editing of at least one variable is applied to 2% of students, while the imputation of at least one variable is applied to only a further 0.1% of students. For less than 0.1% of students, RBI does not provide a fully correct set of variables. For half of these students, the set of variables is already fully correct in the TRET, and the integration with RBI would determine the loss of some information. For the other half, the set is not fully correct in the TRET either. How to treat these cases has to be decided. However, since the impact is negligible, in Section 4 it has been decided to impute the information from RBI when available regardless the fully correctness of the information, and use the data from TRET when not available in RBI.

Table 4: Effects of the integration with RBI

	Students with a single record		**Students with a cluster of consistent records***	
Effect	**Freq.**	**%**	**Freq.**	**%**
None	7,966,948	97.69	190,051	97.94
At least 1 editing	174,877	2.15	3,707	1.91
At least 1 imputation	6,131	0.08	174	0.09
To be decided:				
Not fully correct in RBI but fully correct in TRET	3,087	0.04	78	0.04
Not fully correct in RBI and TRET	4,117	0.05	42	0.02
Total	8,155,160	100	194,052	100

Appendix 2

Table 5: Resident individuals (number and %) in RBI and not in TRET by year of birth and age

Year of birth	Age	n	%
2003	18	125.631	19,8
2004	17	73.931	11,7
2005	16	59.335	9,4
2006	15	49.700	7,8
2007	14	39.345	6,2
2008	13	18.234	2,9
2009	12	18.014	2,8
2010	11	17.875	2,8
2011	10	18.218	2,9
2012	9	18.994	3,0
2013	8	17.872	2,8
2014	7	17.694	2,8
2015	6	16.935	2,7
2016	5	39.744	6,3
2017	4	45.004	7,1
2018	3	58.049	9,2
Total		634.575	100
In compulsory school age - Total		292.216	

Appendix 3

Table 6 reports the comparison between TRET and MIM estimates for the upper secondary school type, which is a relevant variable in tracking the scholastic paths. The goodness of agreement of the estimates is confirmed.

Table 6: Number of attending students (n) estimated from the TRET and percent relative variations (r.v.) respect to MIM open data, for upper secondary school type

Upper secondary school type	n	r.v.
Lyceum	1,367,568	0.0
Technical Institute	838,078	-0,2
Professional Institute	456,181	-0,2