A Unified Education Dataset at the Economics of Education Knowledge Center

João Firmino [1] Gonçalo Lima [2] Luís Nunes [3] Pedro Freitas [4]

[1, 2, 3, 4]Nova SBE

[2]European University Institute

EEKC Seminar

March 4, 2022

NOVA Economics of Education Knowledge Center
NOVA SCHOOL OF BUSINESS & ECONOMICS

FCT Fundação para a Ciência e a Tecnologia

SOCIAL SCIENCES DATALAB

Table of Contents

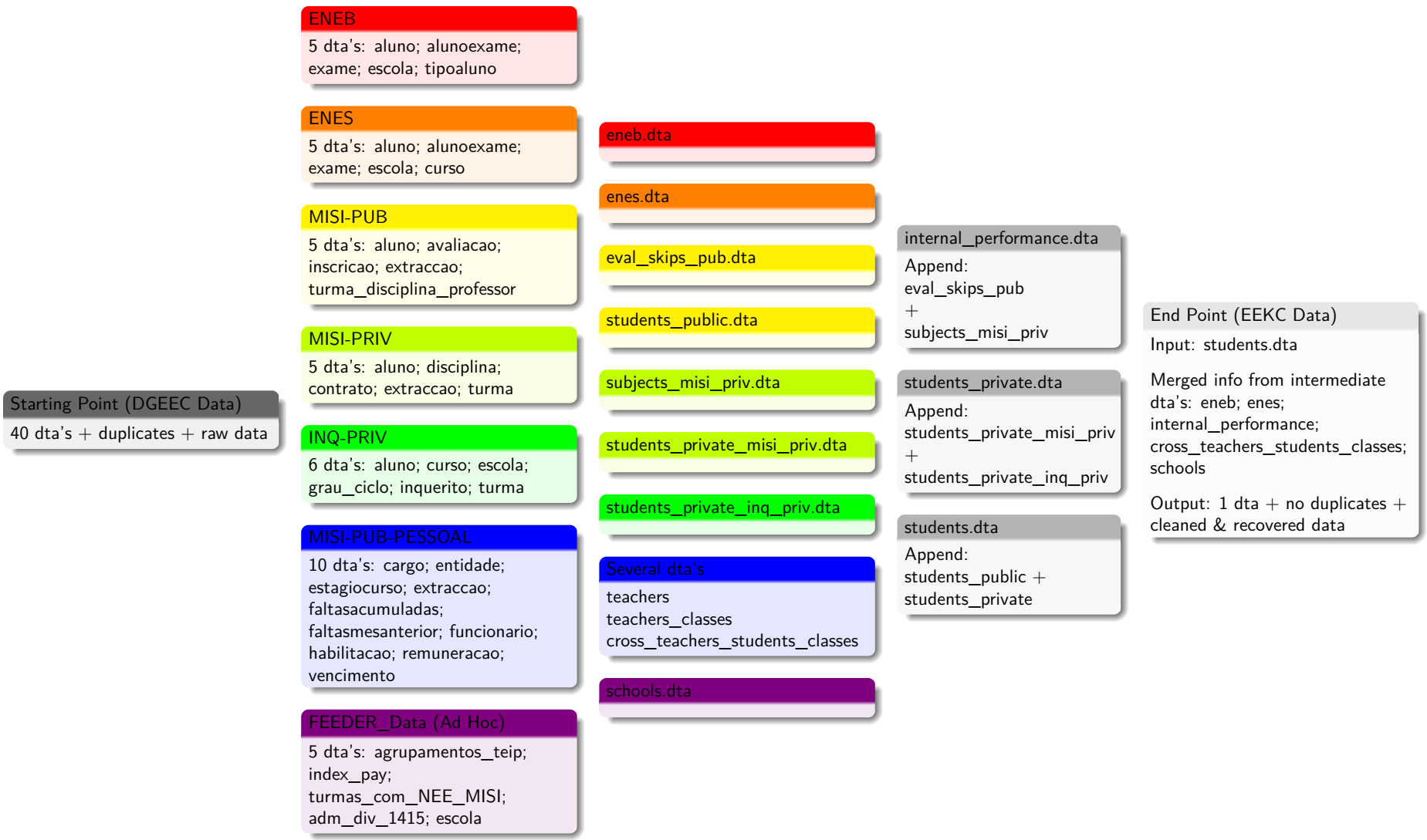
Why a Common Unified Dataset

- Save time!
  - Complete ready-to-use dataset for all users
    - Masters; PhDs; RAs; Faculty; Visiting Researchers
- Publicly available Stata code
  - Transparent assumptions on:
    - Raw datasets' merges/reshapes
    - Duplicates cleaning
    - Cleaning of variables
  - Harmonization of procedures
- In English (future non-Portuguese users)
  - variables: name, label, value label
  - do-file: assumptions, comments
- Now includes 2017/18

Data Sources

- JNE-National Exams Jury:
  - ENEB and ENES: Students' scores on national exams
- DGEEC-Directorate-General for Statistics of Education and Science:
  - MISI-PUB-PESSOAL: Teachers in public schools
  - MISI-PUB: Students in public schools
  - MISI-PRIV: Students in Government dependent private schools
  - INQ-PRIV: Students in Independent private schools

The Map


Starting Point (DGEEC Data)
40 dta's + duplicates + raw data
ENEB
5 dta's: aluno; alunoexame; exame; escola; tipoaluno
ENES
5 dta's: aluno; alunoexame; exame; escola; curso
MISI-PUB
5 dta's: aluno; avaliacao; inscricao; extraccao; turma_disciplina_professor
MISI-PRIV
5 dta's: aluno; disciplina; contrato; extraccao; turma
INQ-PRIV
6 dta's: aluno; curso; escola; grau_ciclo; inquerito; turma
MISI-PUB-PESSOAL
10 dta's: cargo; entidade; estagiocurso; extraccao; faltasacumuladas; faltasmesanterior; funcionario; habilitacao; remuneracao; vencimento
FEEDER_Data (Ad Hoc)
5 dta's: agrupamentos_teip; index_pay; turmas_com_NEE_MISI; adm_div_1415; escola
eneb.dta
enes.dta
eval_skips_pub.dta
students_public.dta
subjects_misi_priv.dta
students_private_misi_priv.dta
students_private_inq_priv.dta
Several dta's
teachers
teachers_classes
cross_teachers_students_classes
schools.dta
internal_performance.dta
Append:
eval_skips_pub
+
subjects_misi_priv
students_private.dta
Append:
students_private_misi_priv
+
students_private_inq_priv
students.dta
Append:
students_public +
students_private
End Point (EEKC Data)
Input: students.dta
Merged info from intermediate dta's: eneb; enes; internal_performance; cross_teachers_students_classes; schools
Output: 1 dta + no duplicates + cleaned & recovered data

ENEB

ENEB

5 dta's: aluno; alunoexame; exame; escola; tipoaluno

Issues & Solutions

Phase duplicates drops:

1. different from student's modal gender & birthdate
2. different from student-by-year modal school
3. missing score (but another row with non-missing score)
4. score lower than max score

Reshape: exams & phases info to columns

ENEB

eneb.dta - phase duplicates

School-Year	% Duplicated	Total (thousands)
2006/07	0.01	326
2007/08	0.01	325
2008/09	0.01	320
2009/10	0.01	321
2010/11	0.01	316
2011/12	0.01	317
2012/13	0.01	319
2013/14	0.01	309
2014/15	0.01	302
2015/16	0.01	93
2016/17	0.01	94
2017/18	0.01	96
All	0.01	3.1M

ENEB

eneb.dta

- obs unique at **student-year** level
- national exams
- low & high stakes
- grades 4, 6, 9
- subjects: Pt, Math

- phase 1 & 2 info:
  - score (final)
  - score (original)
  - score rectified
  - score reevaluated (2nd grader)
  - score reevaluated (3rd grader)

ENES

ENES

5 dta's: aluno; alunoexame; exame; escola; curso

Issues & Solutions

Many exams - kept 10 subjects' exams
Phase duplicates drops:

1. different from student's modal gender & birthdate
2. missing score (but another row with non-missing score)
3. score lower than max score
4. at random

Reshape: exams & phases info to columns

ENES

enes.dta - phase duplicates

School-Year	% Duplicated	Total (thousands)
2006/07	0.00	139
2007/08	0.00	143
2008/09	0.00	146
2009/10	0.00	150
2010/11	0.00	150
2011/12	0.00	151
2012/13	0.00	149
2013/14	0.01	149
2014/15	0.02	149
2015/16	0.02	153
2016/17	0.06	155
2017/18	0.02	151
All	0.01	1.8M

ENES

enes.dta

- national exams
- high stakes
- grades 11, 12
- subjects: Pt, Math, History, Bio&Geol, Draw, Geom, Econ, Phil, Phys&Chem, Geog
- phase 1 & 2 info:
  - score (final)
  - score (original)
  - score rectified
  - score reevaluated (2nd grader)
  - score reevaluated (3rd grader)
  - if student final GPA uses teacher scores [Interno]
  - if exam score used for approval
  - if exam score used to improve past score of already approved subject [Melhoria]
  - if exam score used for univ. application
- obs unique at **student-year** level

MISI-PUB

MISI-PUB (eval_skips_pub.dta)

5 dta's: aluno; avaliacao; inscricao; extraccao; turma_disciplina_professor

Issues & Solutions

1. kept only end of year info [P3, FN, EX]
2. spread info on max # of skips across student-year-subject duplicates (assump: year long skips count)
3. drop [EX] rows
4. drop rows of unimportant or unknown subjects (4 digit subjectid with no clear naming)

Student-year-subject-evaluation duplicates:

- dropped at random

Student-year-subject duplicates - dropped if:

1. missing [P3] score (but another row with non-missing [P3] score)
2. all rows missing [P3] (but another with non-missing [FN]) - only for year/grade comb with no high-stakes exams
3. at random

MISI-PUB

eval_skips_pub.dta - evaluation duplicates

School-Year	% Duplicated	Total (millions)
2006/07	0.00	0.03
2007/08	0.02	5.2
2008/09	0.08	5.8
2009/10	0.07	5.9
2010/11	0.10	6.1
2011/12	0.09	6.2
2012/13	0.17	6.4
2013/14	0.20	6.4
2014/15	0.18	6.2
2015/16	0.17	6.0
2016/17	0.11	5.9
2017/18	0.10	5.7
All	0.12	65.8

MISI-PUB

eval_skips_pub.dta - subject duplicates

School-Year	% Duplicated	Total (millions)
2006/07	52	0.03
2007/08	65	5.2
2008/09	69	5.8
2009/10	70	5.9
2010/11	97	6.1
2011/12	99	6.2
2012/13	98	6.4
2013/14	99	6.4
2014/15	99	6.2
2015/16	96	6.0
2016/17	92	5.9
2017/18	87	5.7
All	89	65.8

MISI-PUB

eval_skips_pub.dta

- scores issued by teacher (end of year)
- student's recorded skips (end of year; just'd & unjust'd)
  - total # of skips per year (across all subjects)
  - average # of skips per year (across all subjects)
- grades 1-12
- ∼30 subjects' info on scores and skips:
  - Pt, Math
  - History, Bio&Geol, Draw, Geom, Econ, Phil, Phys&Chem, Geog, Physical Educ, Eng, Fr, Spa, Ger, Nat Sciences, Mor & Rel Educ, Draw & Tec Educ, Music Educ, Sign Language, Pt for Foreigners
  - recovered 5-digit subjectid when 4-digit id used (4-digit id not on official doc) via subject name
  - new subject name var grouping similar subjects
    - "math a" stands for Basic "math" and Secondary "math a" subjectid's
    - "english" stands for "english", "english i", "english ii", ... subjectid's
- obs unique at **student-year-subject** level

MISI-PUB

eval_skips_pub.dta - new info:

- students with numeric PT score between 2006/07 and 2016/17
- 1.4M extra obs

	Total (millions)	% With Info	With Info (millions)
v1	6.8	77	5.3
v2	7.4	90	6.7

MISI-PUB

MISI-PUB (students_public.dta)

5 dta's: aluno; avaliacao; inscricao; extraccao; turma_disciplina_professor

Issues & Solutions

Student-year-school-class duplicates:

- kept duplicates with most sociodemographic non-missing info

Student-year-school duplicates:

1. kept duplicates with principal = 1
2. kept duplicates with most sociodemographic non-missing info

Student-year duplicates (mostly, school transfers within school-year):

1. kept duplicates with most sociodemographic non-missing info when there is no info about events' dates
   - "events": conclusion of grade, retention, school transfer, ...
2. dropped duplicates referring to "administrative" transfers (happening before school year start)
3. kept duplicates referring to grade conclusion events when at least 1 of these events (or transfers) has no date.
4. kept duplicates referring to school of longest spell
5. dropped duplicates that have different school ID than school ID of longest spell
6. kept duplicates with most sociodemographic non-missing info
7. dropped duplicates at random

MISI-PUB

students_public.dta - duplicates (class & school & year)

School-Year	% Duplicated	Total (millions)
2006/07	3.1	1.2
2007/08	2.8	1.2
2008/09	2.8	1.2
2009/10	3.0	1.2
2010/11	2.9	1.2
2011/12	3.1	1.2
2012/13	3.0	1.1
2013/14	3.1	1.1
2014/15	2.6	1.2
2015/16	2.4	1.2
2016/17	2.0	1.2
2017/18	2.0	1.1
All	2.7	14.2

MISI-PUB

students_public.dta - duplicates (class & school & year)

Curriculum	% Duplicated	Total (thousands)
Kindergarten	1.6	504
Basic-Regular	2.0	10.2M
Basic-CEF	5.4	226
Basic-Adult (EFA)	0.9	81
Basic-Vocational	3.9	47
Basic-Artistic	35.3	21
Basic-PIEF	11.7	24
Basic-Adult (Recorrente)	5.7	12
Secondary-Regular Scientific	5.0	2.0M
Secondary-Regular General	13.4	15
Secondary-Regular Technological	5.2	92
Secondary-CEF	2.5	11
Secondary-Adult (EFA)	1.3	139
Secondary-Vocational	1.6	9
Secondary-Artistic	13.0	31
Secondary-Adult (Recorrente)	12.9	138
Secondary-Professional	2.2	679
All	2.7	14.2M

MISI-PUB

students_public.dta

- all public school students
- simple cleaning/correction/labeling of students & parents' sociodemographics
  - corrected parents' job IDs from CNP1994 to CPP2010 (06/07 to 10/11)
- created var "curriculum" combining:
  - cycle of studies: Kindergarten, ..., Secondary
  - type of education: Regular, Voc, Prof, Adult, ...
- created var "course" with course IDs
  - complements "curriculum" - useful for non-Regular tracks
- obs unique at **student-year** level

MISI-PRIV

MISI-PRIV (subjects_misi_priv.dta)

5 dta's: aluno; disciplina; contrato; extraccao; turma

Issues & Solutions

1. drop rows of unimportant subjects
2. keep rows for which subject info matches with studentid and year info

Student-year-subject duplicates:

1. dropped student-year panels with 2 or more distinct counts of # of subjects
   - only makes sense to have 1 count in a given year
2. dropped at random duplicates due to different subject ids within same subject name

MISI-PRIV

subjects_misi_priv.dta - subject duplicates

School-Year	% Duplicated	Total (thousands)
2006/07		
2007/08		
2008/09	0.26	468
2009/10	0.05	583
2010/11	0.05	500
2011/12	0.05	507
2012/13	0.03	482
2013/14	0.05	502
2014/15	0.08	519
2015/16	0.02	514
2016/17	0.02	456
2017/18	0.17	292
All	0.07	4.8M

MISI-PRIV

subjects_misi_priv.dta

- grades 1-12
- ∼30 subjects' enrollment info:
  - Pt, Math, History, Bio&Geol, Draw, Geom, Econ, Phil, Phys&Chem, Geog
  - Physical Educ, Eng, Fre, Spa, Ger, Nat Sciences, Mor & Rel Educ, Draw & Tec Educ, Music Educ, Sign Language, Pt for Foreigners
- recovered 5-digit subjectid for first dataset years (4-digit id not on official doc)
- new subject name var grouping similar subjects
  - "math a" stands for Basic "math" and Secondary "math a" subjectid's
  - "english" stands for "english", "english i", "english ii", ... subjectid's
- obs unique at **student-year-subject** level

Internal Performance

internal_performance.dta

Append: eval_skips_pub + subjects_misi_priv

Issues & Solutions

Students exported twice: via MISI-PUB & via MISI-PRIV
Student-year-source duplicates:

- dropped duplicates exported via misi-priv (no students' evaluation/skips info)

Internal Performance

internal_performance.dta - source duplicates

School-Year	% Duplicated	Total (millions)
2006/07	N/A	33k
2007/08	N/A	5.2
2008/09	0.11	6.3
2009/10	0.16	6.4
2010/11	0.27	6.6
2011/12	0.27	6.7
2012/13	0.27	6.9
2013/14	0.31	6.8
2014/15	0.29	6.7
2015/16	0.27	6.5
2016/17	0.22	6.3
2017/18	0.16	6.0
All	0.22	70.5

Internal Performance

internal_performance.dta

- ∼30 subjects' info:
  - Pt, Math, History, Bio&Geol, Draw, Geom, Econ, Phil, Phys&Chem, Geog
  - Physical Educ, Eng, Fre, Spa, Ger, Nat Sciences, Mor & Rel Educ, Draw & Tec Educ, Music Educ, Sign Language, Pt for Foreigners
- info on subject enrollment: MISI-PUB & MISI-PRIV
- info on teacher scores & students' skips: MISI-PUB
- obs unique at **student-year-subject** level

MISI-PRIV

MISI-PRIV (students_private_misi_priv.dta)

5 dta's: aluno; disciplina; contrato; extraccao; turma

Issues & Solutions

Student-year-school-class duplicates:

- kept duplicates with most sociodemographic non-missing info

Student-year-school duplicates:

1. kept duplicates with principal = 1
2. drop duplicates whose grade as stated in class name disagrees with var "grade"
3. kept duplicates with most sociodemographic non-missing info
4. dropped at random

Student-year duplicates (mostly, school transfers within school-year):

1. kept duplicates with most sociodemographic non-missing info when there is no info about events' dates
   - "events": conclusion of grade, retention, school transfer, ...
2. dropped duplicates referring to "administrative" transfers (happening before school year start)
3. kept duplicates referring to grade conclusion events when at least 1 of these events (or transfers) has no date.
4. kept duplicates referring to school of longest spell
5. dropped duplicates that have different school ID than school ID of longest spell
6. kept duplicates with most sociodemographic non-missing info
7. dropped at random

MISI-PRIV

students_private_misi_priv.dta - duplicates (class & school & year)

School-Year	% Duplicated	Total (thousands)
2006/07		
2007/08		
2008/09	0.7	67
2009/10	1.6	82
2010/11	1.8	76
2011/12	0.8	81
2012/13	0.9	77
2013/14	0.9	82
2014/15	0.8	83
2015/16	0.8	86
2016/17	0.5	77
2017/18	0.6	53
All	1.0	763

MISI-PRIV

students_private_misi_priv.dta - duplicates (class & school & year)

Curriculum	**% Duplicated**	**Total (thousands)**
Kindergarten	0.1	13
Basic-Regular	0.2	422
Basic-CEF	1.0	16
Secondary-Regular Scientific	2.7	113
Secondary-Regular Technological	1.0	4
Secondary-CEF	4.1	2
Secondary-Adult (EFA)	0.4	3
Secondary-Professional	1.7	162
Secondary-Specific Plans	1.3	27
All	1.0	763

MISI-PRIV

students_private_misi_priv.dta

- private school students with State subsidized fees
- Simple cleaning/correction/labeling of students' sociodemographics
  - recovered MISI-PRIV students' district and municipality from parish
- created var "curriculum" combining:
  - cycle of studies: Kindergarten, ..., Secondary
  - type of education: Regular, Voc, Prof, Adult, ...
- created var "course" with course IDs
- created var "fund_type" with type of State funding given to student to enroll in private school
- obs unique at **student-year** level

INQ-PRIV

INQ-PRIV

5 dta's: aluno; curso; escola; grau_ciclo; inquerito; turma

Issues & Solutions

Student-year-school-class duplicates:

- kept duplicates with most sociodemographic non-missing info

Student-year-school duplicates:

- kept duplicates with most sociodemographic non-missing info

Student-year duplicates (mostly, school transfers within school-year):

1. kept duplicates referring to full-time learning (provided both part-time and full-time duplicates exist)
2. dropped duplicates referring to transfer events (when at least 1 grade conclusion event is observed)
   - "events": conclusion of grade, retention, school transfer, ...
3. kept duplicates with most sociodemographic non-missing info

INQ-PRIV

students_private_inq_priv.dta - duplicates (class & school & year)

School-Year	% Duplicated	Total (thousands)
2006/07	0.3	167
2007/08	1.3	195
2008/09	0.4	159
2009/10	0.3	138
2010/11	0.3	130
2011/12	0.3	129
2012/13	0.3	127
2013/14	0.8	124
2014/15	0.4	125
2015/16	0.3	128
2016/17	0.4	130
2017/18	0.3	240
All	0.5	1.8M

INQ-PRIV

students_private_inq_priv.dta - duplicates (class & school & year)

Curriculum	% Duplicated	Total (thousands)
Kindergarten	0.2	108
Basic-Regular	0.3	941
Basic-CEF	0.6	10
Basic-Vocational	0.9	10
Basic-Professional	1.5	6
Secondary-Regular Scientific	0.7	151
Secondary-Adult (Recorrente)	0.6	25
Secondary-Professional	0.6	400
Secondary-Specific Plans	13.1	9
Basic	0.1	64
Secondary	0.1	14
Adult (Recorrente)	0.8	17
CET	0.9	12
Unknown	0.3	12
Other		
All	0.5	1.8M

INQ-PRIV

students_private_inq_priv.dta

- private school students without State subsidized fees
- Simple cleaning/correction/labeling of students' sociodemographics
  - recovered INQ-PRIV students' district and municipality from postal code
- created var "curriculum" combining:
  - cycle of studies: Kindergarten, ..., Secondary
  - type of education: Regular, Voc, Prof, Adult, ...
- created var "course" with course IDs
- corrected names of mixed grade classes
  - schools mandated to export info on mixed grade classes by dividing them in pseudo-single-grade-classes
  - problem: more single grade classes & fewer mixed grade classes & smaller class sizes
    - no correction: 14 obs under mixed grade class
    - correction: 11k obs under mixed grade class
    - out of 470k obs (grades 1 to 4, all years up to 1617)
- obs unique at **student-year** level

Private Schools Students

students_private.dta

Append: students_private_misi_priv + students_private_inq_priv

Issues & Solutions

Students exported twice: via MISI-PRIV & via INQPRIV

Student-year-school-class duplicates:

1. kept largest "duplicated" class (to preserve info of non-duplicated students)
2. kept duplicated class exported via misi-priv (more students' characteristics and funding type info)

Student-year-school duplicates:

1. dropped classes where half (or more) of classmates are indentified as duplicates and class exported via inqpriv (assumes such classes have corresponding duplicated classes exported via misi-priv but with a different name - hence not captured in previous steps; misi-priv classes with more info)
2. kept duplicates with principal = 1
3. kept duplicates of highest grade (if duplications report different grades; clas of highest grade should be principal class)
4. dropped at random

Student-year duplicates:

- kept duplicates exported via misi-priv (more info about student)

Private Schools Students

students_private.dta - duplicates (class & school & year)

School-Year	% Duplicated	Total (thousands)
2006/07	0.3	168
2007/08	1.3	195
2008/09	10.0	206
2009/10	5.4	210
2010/11	1.0	206
2011/12	0.6	210
2012/13	0.9	203
2013/14	1.1	205
2014/15	0.8	208
2015/16	1.2	212
2016/17	0.7	207
2017/18	0.5	293
All	2.0	2.5M

Private Schools Students

students_private.dta - duplicates (class & school & year)

Curriculum	% Duplicated	Total (thousands)
Kindergarten	0.2	121
Basic-Regular	1.5	1.3M
Basic-CEF	3.5	26
Basic-Vocational	0.9	10
Secondary-Regular Scientific	3.3	260
Secondary-Regular Technological	6.0	4
Secondary-CEF	30.8	2
Secondary-Adult (Recorrente)	0.8	26
Secondary-Professional	2.5	551
Secondary-Specific Plans	10.5	34
Basic	0.1	64
Secondary	0.1	14
Adult (Recorrente)	0.8	16
CET	1.1	12
Unknown	0.3	12
Other		
All	2.0	2.5M

Private Schools Students

students_private.dta

- all private school students
- recovered info on students under simple contract:
  - MISI-PRIV: fund_type == "Simple Contract"
  - INQPRIV: Simple Contract if $1 \leq$ echelon ≤ 4
    - before info recovery: in 15/16, 8k students under simple contract (only misi-priv info)
    - after info recovery: 25k students under simple contract (misi-priv + inqpriv info)
    - Expresso article states that in 15/16 the Ministry reported 22k such students
- obs unique at **student-year** level

All Students

students.dta

Append: students_public + students_private

Issues & Solutions

Student-year duplicates (school transfers within schoolyear between public and private schools)

1. kept duplicates referring to grade conclusion events (discard transfer/drop-out events)
2. kept duplicate with most sociodemographic info

All Students

students.dta - duplicates (pub/priv within schoolyear)

School-Year	% Duplicated	Total (millions)
2006/07	0.4	1.4
2007/08	0.5	1.4
2008/09	0.6	1.4
2009/10	0.6	1.4
2010/11	0.6	1.4
2011/12	0.6	1.4
2012/13	0.6	1.3
2013/14	0.8	1.3
2014/15	0.6	1.5
2015/16	0.6	1.4
2016/17	0.5	1.4
2017/18	0.7	1.4
All	0.6	16.9

All Students

students.dta - duplicates (pub/priv within schoolyear)

Curriculum	% Duplicated	Total (thousands)
Kindergarten	0.5	621
Basic-Regular	0.3	11.5M
Basic-CEF	2.3	248
Basic-Adult (EFA)	0.2	82
Basic-Vocational	5.2	56
Basic-Artistic	19.1	24
Basic-PIEF	0.9	24
Basic-Adult (Recorrente)	2.8	15
Secondary-Regular Scientific	0.6	2.2M
Secondary-Regular General	0.2	15
Secondary-Regular Technological	0.4	95
Secondary-CEF	1.3	12
Secondary-Adult (EFA)	0.6	141
Secondary-Artistic	2.1	33
Secondary-Adult (Recorrente)	1.5	163
Secondary-Professional	2.4	1.2M
CET	0.3	12
Other		
All	0.6	16.9M

All Students

students.dta - duplicates (class & school & year)

School-Year	% Duplicated	Total (millions)
2006/07	2.7	1.4
2007/08	2.6	1.4
2008/09	3.7	1.4
2009/10	3.2	1.4
2010/11	2.5	1.4
2011/12	2.7	1.4
2012/13	2.6	1.4
2013/14	2.8	1.3
2014/15	2.3	1.5
2015/16	2.2	1.4
2016/17	1.8	1.4
2017/18	1.7	1.4
All	2.6	16.9

All Students

students.dta - duplicates (class & school & year)

Curriculum	% Duplicated	Total (thousands)
Kindergarten	1.3	621
Basic-Regular	2.0	11.5M
Basic-CEF	5.2	248
Basic-Adult (EFA)	0.9	82
Basic-Vocational	3.4	56
Basic-Artistic	30.4	24
Basic-PIEF	11.7	24
Basic-Adult (Recorrente)	4.8	15
Secondary-Regular Scientific	4.8	2.2M
Secondary-Regular General	13.3	15
Secondary-Regular Technological	5.3	95
Secondary-CEF	5.8	12
Secondary-Adult (EFA)	1.3	141
Secondary-Artistic	12.3	33
Secondary-Adult (Recorrente)	11.0	163
Secondary-Professional	2.3	1.2M
CET	1.1	12
Other		
All	2.6	16.9M

All Students

students.dta - sources (thousands; millions (M))

School-Year	MISI-PUB	MISI-PRIV	INQPRIV	ENEB	ENES	Total
2006/07	1.2M	0	166	31	10	1.4M
2007/08	1.2M	0	193	25	7	1.4M
2008/09	1.2M	65	138	22	7	1.4M
2009/10	1.2M	80	126	22	8	1.4M
2010/11	1.2M	75	128	24	8	1.4M
2011/12	1.2M	80	127	31	8	1.4M
2012/13	1.1M	75	125	24	8	1.4M
2013/14	1.1M	80	123	21	7	1.3M
2014/15	1.2M	81	124	21	8	1.5M
2015/16	1.2M	84	125	6	8	1.4M
2016/17	1.2M	76	129	6	8	1.4M
2017/18	1.1M	52	239	0	13	1.4M
All	14.1M	746	1.7M	233	101	16.9M

All Students

Issues & Solutions

Missing/Inconsistent Data

1. recover student info from ENEB and ENES
   - gender; birthdate; socioeconomic status (SASE); parent education
2. within student panel corrections:
   - **mode**: gender; birthdate; birthplace (student's and parents')
   - **interpolation only**: replace missing by interpolated value if missing is within a gap of X schoolyears with known values
     - $X = 1$: computer; internet & alternative curriculum; student's domicile (parish; municipality)
     - $X = 3$: family support (Abono Familia)
     - $X = 5$: student's guardian kinship
     - $X = \infty$: student's SEN status
   - **interpolation/extrapolation**: given known data-point, use it to replace missings up to X periods apart (nearest neighbor)
     - $X = 4$: parent occupation (4 digit ID); parent job status
     - $X = \infty$: parent education

All Students

students.dta - recovered data: mother place of birth (% with info)

School-Year	MISI-PUB	MISI-PRIV	INQPRIV	ENEB	ENES
2006/07	94 → 95		0 → 32	0 → 4	0 → 4
2007/08	96 → 97		0 → 42	0 → 3	0 → 3
2008/09	96 → 97	0 → 49	0 → 52	0 → 2	0 → 4
2009/10	96 → 98	0 → 59	0 → 57	0 → 3	0 → 6
2010/11	96 → 98	0 → 66	0 → 56	0 → 4	0 → 7
2011/12	97 → 99	0 → 70	0 → 55	0 → 5	0 → 9
2012/13	96 → 99	0 → 75	0 → 53	0 → 11	0 → 9
2013/14	96 → 99	0 → 77	0 → 53	0 → 7	0 → 7
2014/15	98 → 99	0 → 78	0 → 51	0 → 6	0 → 8
2015/16	98 → 100	0 → 77	0 → 48	0 → 8	0 → 9
2016/17	99 → 100	0 → 76	0 → 43	0 → 10	0 → 10
2017/18	99 → 100	0 → 78	0 → 22		0 → 39

All Students

students.dta - recovered data: mother education (% with info)
(All cohorts)

School-Year	MISI-PUB	MISI-PRIV	INQPRIV	ENEB	ENES
2006/07	79 → 86		0 → 25	0 → 3	0 → 2
2007/08	80 → 89		0 → 34	0 → 2	0 → 2
2008/09	79 → 89	0 → 40	0 → 45	0 → 2	0 → 3
2009/10	77 → 90	0 → 51	0 → 51	0 → 2	0 → 5
2010/11	77 → 90	0 → 58	0 → 51	0 → 3	0 → 6
2011/12	78 → 91	0 → 63	0 → 50	0 → 6	0 → 8
2012/13	79 → 92	0 → 71	0 → 49	0 → 17	0 → 6
2013/14	79 → 92	0 → 71	0 → 49	0 → 17	0 → 6
2014/15	83 → 93	0 → 73	0 → 47	0 → 7	0 → 7
2015/16	83 → 93	0 → 73	0 → 45	0 → 12	0 → 11
2016/17	84 → 92	10 → 73	3 → 42	52 → 56	8 → 17
2017/18	85 → 92	11 → 75	2 → 22		11 → 42

All Students

students.dta - recovered data: mother education (% with info)
(Cohorts in grade 9 by 2016/17 or 2017/18: parent education info via ENEB)

	MISI-PUB	MISI-PRIV	INQPRIV
2006/07	77 → 93		0 → 76
N	10,103		403
2007/08	76 → 93		0 → 78
N	28,373		1,289
2008/09	82 → 95	0 → 74	0 → 78
N	98,038	1,266	10,225
2009/10	83 → 95	0 → 76	0 → 75
N	156,207	1,820	18,551
2010/11	84 → 95	0 → 75	0 → 74
N	157,345	1,922	18,648
2011/12	84 → 95	0 → 81	0 → 74
N	158,293	3,190	17,675
2012/13	84 → 94	0 → 90	0 → 71
N	158,329	8,063	14,266
2013/14	84 → 94	0 → 90	0 → 69
N	158,202	12,909	11,405
2014/15	85 → 94	0 → 90	0 → 68
N	159,120	13,769	10,691
2015/16	85 → 94	0 → 90	0 → 67
N	160,576	14,240	10,051
2016/17	86 → 93	37 → 90	24 → 67
N	162,773	13,593	10,458
2017/18	83 → 93	35 → 90	17 → 76
N	154,218	11,624	15,139

All Students

students.dta

- how to identify classes:
  - for merges with raw datasets (but NOT for analyses)
    - variable "class": original "turma" variable
  - for analyses (but NOT for merges with raw datasets)
    - variable "class2": original "turma" variable corrected for mixed grade classes
    - variable "class_eekc": class ID unique for any year-school_eekc-class2 combination
- how to identify schools:
  - for merges with raw datasets and analyses
    - variable "school": original "escola" variable
    - variable "school_dgeec": DGEEC school ID (may contain different "school/escola" ids hence more constant than "school/escola")
  - for analyses (but NOT for merges with raw datasets)
    - variable "school_eekc": school ID unique for any DGEEC ID; then, if DGEEC ID missing, unique for "school/escola" ID; then, if both DGEEC and school/escola IDs missing, unique for any school name-school location combination

All Students

students.dta

- info on all students (public + private) with panel corrections
- merged info about SEN/Limited Portuguese Proficiency
  - via national exams (ENEB & ENES)
    - SEN: "NE - Nível Escola" exams
    - LPP: "PLNM" exams
  - via subjects (MISI-PUB & MISI-PRIV)
    - SEN: Sign language subject
    - LPP: "PLNM" subject
- obs unique at **student-year** level

MISI-PUB-PESSOAL

MISI-PUB-PESSOAL

10 dta's: cargo; entidade; estagiocurso; extraccao; faltasacumuladas; faltasmesanterior; funcionario; habilitacao; remuneracao; vencimento; funcionario

FEEDER_Data (Ad Hoc)

5 dta's: agrupamentos_teip; **index_pay**; turmas_com_NEE_MISI; adm_div_1415; schools_plus

teachers

- **duplicates**: teacher-schoolyear (∼monthly extractions)
  1. dropped duplicates if flagged as non-teachers
  2. kept first duplication after taking mode within teacher-schoolyear duplication panel across relevant variables

MISI-PUB-PESSOAL

teachers

- experience (corrected for 0s & negative/too large yearly changes)
- wage (components: base pay, lunch subsidy, ...)
- skips (types: own health, relatives health, ...)
- education (completed, university final GPA)
- internship (type, final GPA, conclusion date, ...)
- contract (begin/end dates, ending reason, substitute teacher, ...)
- career stage (10 echelons/payment indexes)
- gross, net, & real monthly wage per stage & year
- order of panel corrections:
  1. mode across same teacher-by-year duplications
  2. drop of same teacher-by-year duplications
  3. mode/interpolation within teacher panel
- obs unique at **teacher-year** level

MISI-PUB-PESSOAL

MISI-PUB

1 dta: **turma_disciplina_professor**

turma_disciplina_professor

- provides class-subject-teacher link
- **duplicates**: teacher-subject-class-school-schoolyear
  - drops:
    1. if 0 time-slots;
    2. if time-slots $<$ max time-slots;
    3. at random
  - dup_teacher_w_subject_class variable:
    - 0.1% to 0.4% of PT classes affected
- recovered 5-digit subjectid for first dataset years (4-digit id not on official doc)

MISI-PUB-PESSOAL

teachers_classes.dta

teachers.dta + turma_disciplina_professor.dta

Issues & Solutions

duplicates: subject-class-school-schoolyear

- 2 or more teachers teaching same subject to same class
- kept:
  1. teachers with teaching roles
  2. substitute teachers whose substitution contract lasts more than half schoolyear
  3. teachers with fewest skips
  4. teachers with not too large base pay drop
  5. at random
- dup_subject_w_class variable:
  - 2006/07: 5% of PT classes affected
  - 2017/18: 13% of PT classes affected

MISI-PUB-PESSOAL

teachers_classes.dta

- teachers linked to classes (thus to students)
- teacher info on:
  - subject lectured
  - skips
  - sociodemographics (panel corrected)
  - education; internship
  - contract; wage
  - experience (panel corrected)
- only public schools
- obs unique at **year-school-class-subject** level

MISI-PUB-PESSOAL

cross_teachers_classes_students

- teachers_classes.dta reshaped:
  - ∼30 subjects moved to columns
    - so that teacher id merges into students.dta preserving idea of 1 row-1 year-student obs (matched to a single school, to a single class, and to a single teacher in subject s)
  - info on:
    - year
    - school
    - class
    - teacher id
  - obs unique at **year-school-class** level

Schools

FEEDER_Data (Ad Hoc)

5 dta's: **agrupamentos_teip**; index_pay; turmas_com_NEE_MISI; adm_div_1415; **escola**

schools

- Input: escola.dta [DGEEC]
  - School info:
    - IDs (escola; coddgeec) & official name
    - location (zip code, parish, municipality, district, GPS coordinates)
    - educational stages offered (kindergarten, basic, ...)
    - public/private
    - tutelage (Ministry of Education, Ministry of Justice, ...)
  - Clusters' info:
    - IDs (escola; coddgeec) & official name
- Output: schools.dta
  - merge info whether school in TEIP program [agrupamentos_teip]
  - appended schools observed across several places not already in escola.dta:
    - ENES, ENEB, MISI-PUB (excel), MISI-PUB-PESSOAL (excel), MISI-PRIV (dta & excel), INQPRIV
  - created EEKC school ID (all schools have at least 1 identifying element: name, escola ID, DGEEC ID; but many schools have missing in 1 element; the EEKC school ID is then filled to all students, public & private)

Main Dataset

main_dataset

- Input: **students.dta**
  (to which the following info has been merged/created)
  - national exams' scores (Pt & Math)
  - teachers' scores & students' skips (Pt & Math)
  - teachers' IDs (Pt & Math)
    - substituted missing Pt teacher ID by Math teacher ID (and vice-versa) for primary schooling (must be same teacher)
  - school info
  - class info:
    - class level modes: data source; student (integer) age; grade; curriculum; course; SEN status; LPP status
    - class level counts: data source; grade; curriculum; course; SEN status; LPP status
    - created EEKC class ID (classes were uniquely identifiable only via the schoolyear-school-class name combination)
    - whether class contains at least 1 SEN student [turmas_com_NEE_MISI]
- Output: **main_dataset.dta**

Future Steps

- to improve:
  - how student-year duplicates drop is handled:
    - identification of longest spell within schoolyear assumes students don't transfer between public/private schools within schoolyear... strong assumption
  - how student demographics from ENEB and ENES is stored:
    - cleaner if code dealing with ENEB/ENES were divided in 2 distinct parts: one devoted to national exams (as is already done), and another entirely devoted to store students' demographics for later use in the panel corrections part (as done in students_eneb and students_enes dta's); right now a bit ad-hoc via preserve/restore shells located within the former part...
  - how to recover INQPRIV students' info on course:
    - use classes exported twice - via MISI-PRIV & via INQPRIV - to understand which courses all INQPRIV students were enrolled in (from INQPRIV side there are course ids matched to course names, but from MISI-PRIV side there are such names)
    - use students that during secondary, up to a certain schoolyear, have info on course name and then exptrapolate which "fkcurso" it corresponds to for schoolyears where those names are no longer provided
  - introduce time dimension when creating schools.dta:
    - "escola" IDs may change over time
    - schools' names may change over time
    - schools' assignment to school clusters may change over time

Future Steps

- (re)validate teachers data (someone?)
- codebook/working paper
- keep merging!
  - post-2017/2018 MISI data
  - autonomous regions' MISI-type data
  - Higher Education national contest data
  - RAIDES
  - private tutoring (demand/supply/prices)
  - Quadros de Pessoal
  - youth sports/health/crime data
  - ...

Acknowledgments

- Data provider:
  - DGEEC
    - Luísa C. C. Loura
    - Joana Duarte
- Research assistants:
  - Diogo Pereira (Nova SBE)
- Funding sources:
  - Fundação para a Ciência e a Tecnologia (FCT)
  - Social Sciences DataLab

APPENDIX

Notes

- all panel corrections performed within students/teachers panels such that the original variables are left unchanged and the corrections are saved in other variables with similar naming/labelling
  - allows researcher to choose/compare among corrected/unocrrected variables
- the original nationality variable is left unchanged, but now there are variables recording the 1st and 2nd nationality of students/parents/teachers
  - 1st nationality: panel 1st most frequent nationality
  - 2nd nationality: panel 2nd most frequent nationality

How variables are grouped in main_dataset

- **source**: original source of the observation
- **studentid**: student unique identifier
- **year**: schoolyear
- **grade-cycle_overall**: student grade/cycle related variables
- **class-class_eekc**: class name/identifier related variables
- **main_class**: whether is the main class of student or not
- **modal_source-n_lpp2**: variables on modal/number of distinct class level characteristics
- **school_cluster-school_name**: variables related to school/school cluster name/identifier
- **parish_schl-ownership_schl_enes**: school level info
- **type_edu-echelon**: pub/priv schooling info & type of funding to attend private schooling
- **stay-dropout**: info on student academic situation/school transfers/spell at school
- **curriculum-regime**: types of curriculum/course/education and their characteristics
- **gender2-zip_locality**: students' characteristics
- **guardian2**: who is the guardian of the student
- **nat1_guardian-job_status2_father2**: info on guardian/mother/father of student
- **dup_w_class-dup_w_year_pub_pri**: info on which observations had a duplication solved
- **n_subjects-average_uskips**: info related to entire set of subjects the student was enrolled
- **orig_subjectid_pt-teacherid_math_a**: Pt/Math A school subjects related variables
- **exam_type-score_12_mat_ph1**: national exams related variables

Example 1 - getting other national exams & related info

1. use main_dataset_18_02
2. merge 1:1 studentid year using eneb_18_02, keepusing(orig_score_4_pt_ph1 score_6_mat_ph2)
3. drop if _merge==2
4. drop _merge
5. merge 1:1 studentid year using enes_18_02, keepusing(score_pt_ph2 to_increase_mat_ph1 score_hist_ph1 score_bio_geol_ph2)
6. drop if _merge==2
7. drop _merge
8. rename score_pt_ph2 score_12_pt_ph2

Example 2 - getting other subjects' teacher scores & skips

```
use internal_performance_18_02, clear
keep studentid year orig_subjectid subject module teacher_score jskips uskips
keep if (subject=="phy_edud" | subject=="geom_a")
rename (orig_subjectid module teacher_score jskips uskips) (orig_subjectid_ module_ teacher_score_ jskips_ uskips_)
greshape wide orig_subjectid_ module_ teacher_score_ jskips_ uskips_, i(studentid year) j(subject)
label var orig_subjectid_geom_a "Original Geometry A Subject Code (4-digit first 2 years; 5-digit remaining years)"
label var orig_subjectid_phy_edud "Original Phys Ed Subject Code (4-digit first 2 years; 5-digit remaining years)"
label var module_geom_a "Whether Geometry A divided in modules during school-year (=1 if so, =0 if not)"
label var module_phy_edud "Whether Phys Ed divided in modules during school-year (=1 if so, =0 if not)"
label var teacher_score_geom_a "Geometry A Score: teacher-issued; full school-year achievement; tipo==P3 (mostly))"
label var teacher_score_phy_edud "Phys Ed Score: teacher-issued; full school-year achievement; tipo==P3 (mostly))"
label var jskips_geom_a "Geometry A justified skips throughout the academic year"
label var jskips_phy_edud "Phys Ed justified skips throughout the academic year"
label var uskips_geom_a "Geometry A unjustified skips throughout the academic year"
label var uskips_phy_edud "Phys Ed unjustified skips throughout the academic year"
order studentid year orig_subjectid_phy_edud jskips_phy_edud uskips_phy_edud module_phy_edud teacher_score_phy_edud orig_subjectid_geom_a jskips_geom_a uskips_geom_a module_geom_a teacher_score_geom_a
save int_perf_aux, replace
```

Example 2 - getting other subjects' teacher scores & skips

1. use main_dataset_18_02, clear
2. merge 1:1 studentid year using int_perf_aux
3. drop if _merge==2
4. drop _merge
5. erase int_perf_aux.dta

Example 3 - getting other students' variables

1. use main_dataset_18_02, clear
2. merge 1:1 studentid year using students_18_02, keepusing(birthday educ_mother dup_w_year_pub_priv)
3. drop if _merge==2
4. drop _merge

Example 4 - getting other teachers' ids and birthday

1. use main_dataset_18_02
2. foreach s in eng hist_a {
   . merge m:1 year school class using cross_teachers_classes_students_18_02, keepusing(teacherid_'s')
   . drop if _merge==2
   . drop _merge
   . rename teacherid_'s' teacherid
   . merge m:1 teacherid year using teachers_18_02, keepusing(birthday_teacher)
   . drop if _merge==2
   . drop _merge
   . rename (teacherid birthday_teacher) (teacherid_'s' birthday_teacher_'s')
   }

Example 5 - getting other school info

1. use main_dataset_18_02
2. merge m:1 school_eekc school using schools_18_02, keepusing(lat_schl long_schl teip)
3. drop if _merge==2
4. drop _merge

Translations

main_dataset	data_anon
studentid	parent
year	anoletivo
school	escola
class	turma
grade	anoensino
cycle	ciclo
school cluster	agrupamento
situation	situacaoactual/situacao
jskips	faltas justificadas
uskips	faltas injustificadas
subject	disciplina