Addition of year 2024 to the anonymised learner-level dataset

16 June 2025[1]

Contents

1 Introduction

The current document accompanies the following two data tables, both Stata version 16 files, with the first one containing labelling:

Table 1: Details on the two tables

File name	Size in KB	Variables	Observations
learner-2024-v1	467,567	12	13,679,374
school-2024-v1	153	3	25,542

The two tables are extensions of earlier tables finalised previously, which contain details for the years 2017 to 2023. The extension is thus an extension of one year, and the abovementioned files contain data just for the year 2024. The anonymisation process ensures that anonymised identifiers for 2024 are linkable to those in the previous years. An analyst wanting to track learner movements from 2017 to 2024 would thus need to combine data files from the earlier work, in particular *learner-2017-2021-v1.dta*, *learner-2022-2023-v1.dta* and the abovementioned *learner-2024-v1.dta*.

Importantly, an initial version of the 2017 to 2021 dataset, produced in September 2022, should not be used, as it was discovered that this had certain problems which were fixed in December 2022. The corrected December 2022 version of the data file is the data file available from DataFirst that is labelled version 1, as the first public release version.

The December 2022 version of the data came with a comprehensive technical report dated 15 December 2022 and with the heading 'An anonymised five-year learner-level dataset for 2017 to 2021'. The 2022 to 2023 data come with the technical report 'Addition of years 2022 and 2023 to the anonymised learner-level dataset' version 1. The current report should be read with the earlier reports. The current report attempts to present new information in the same format as before, without unduly repeating the earlier information.

2 The original source data

For the data file *school-2024-v1.dta*, new information was sourced from the 2024 quarter 3 master lists of schools available on the DBE website as ten separate Excel files, one for each the nine provinces, and a national file for special schools. The ten files were downloaded in April 2025.

[1] Produced by Martin Gustafsson (mgustafsson@sun.ac.za) for the Department of Basic Education.

3 Merging and normalisation occurring before anonymisation

Table 2 below compares learners in the received 2024 data to officially reported learners. As in the previous report, 'Other' in the table refers to learners who are either not in grades R to 12, or who are not in an ordinary school. Alignment between the data and official statistics is fairly high. For instance, with respect to grades R to 12 in ordinary schools, public plus independent, in 2024 the total from the data of 13,479,043 is only 0.15% lower than the total from the official reports, the difference being 20,770 learners. Virtually all of this discrepancy relates to grades 8 to 12 in Gauteng. The explanation could lie in the fact that certain Gauteng schools were absent in the 2024 quarter 3 master list of schools – being in an ordinary school for Table 2 was determined by the presence and classification of a school in the master list. The 'Other' in the data would be mostly special school learners, but also a few learners in ordinary schools but not in a regular grade in the range R to 12. In the official enrolment report, the latter accounts for just 1,449 learners. Special school enrolments, which are not reported on in the official national enrolment reports, are said to come around 140,000 currently. The 200,331 'Other' in Table 2 minus the Gauteng secondary discrepancy and minus the abovementioned 1,449 produces around 180,000 learners. If of these, 140,000 are special school learners, that leaves around 40,000 learners unclassified. Closer analysis of the data, in particular school EMIS numbers, is very likely to explain this.

Table 2: Official totals versus totals in the data for 2024

	Official	Data	%
EC	1,790,055	1,789,488	-0.03
FS	715,866	715,866	0.00
GP	2,659,435	2,639,237	-0.76
KN	2,889,094	2,889,092	0.00
LP	1,817,806	1,817,806	0.00
MP	1,160,068	1,160,067	0.00
NC	308,690	308,690	0.00
NW	882,286	882,285	0.00
WC	1,276,513	1,276,512	0.00
Total	13,499,813	13,479,043	-0.15
Other		200,331	
Final	13,499,813	13,679,374	1.33

There are no variables where the data are completely missing in the 2024 data. The following two tables refer to variables of the *pre*-anonymised data, and where values are missing. The situation for 2024 is virtually the same as for 2023. It should be noted that any grade not in the range of 0 (or R) to 12 was converted to missing, even if some value appeared in the original.

Table 3: Percentage missing by year

	2017	2018	2019	2020	2021	2022	2023	2024
Idno	13.3	8.5	7.5	6.6	6.8	7.3	7.3	7.3
accessionno	0.2	0.0	0.0	13.9	0.0	0.0	0.0	0.0
Birthdate	13.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Fname	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0
Sname	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Grade	0.2	0.2	0.3	0.3	0.3	0.3	0.4	0.4
Class	100.0	8.9	100.0	13.9	0.0	0.0	0.0	0.0
Gender	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Race	100.0	9.0	100.0	0.0	0.0	0.0	0.0	0.0

Table 4: Percentage missing by province for 2024

	EC	FS	GP	KN	LP	MP	NC	NW	WC
Idno	4.8	6.2	13.8	5.3	3.8	7.2	2.4	6.3	8.7
accessionno	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Birthdate	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Fname	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sname	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Grade	0.3	0.4	0.8	0.3	0.1	0.2	0.4	0.5	0.4
Class	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Gender	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Race	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

4 How the 2024 data were anonymised

The anonymisation process was relatively straightforward, and followed the methods employed for the earlier datasets. The five anonymised variables took on the earlier anonymised values wherever possible, so if 'KHUMALO' in Sname was anonymised as 49520 in the 2017 to 2021 data, that same 49520 would be used in the new data. Thereafter, values which were not found in the earlier data were translated to new anonymous values. Table 5 provides the details. For instance, in idno_anon the maximum value in the 2017 to 2023 data was 19360783, meaning the new series of identifiers had to start at 19360784, and continued to 20362199, giving 1,001,416 new and unique anonymised 13-digit identity numbers as this was the number of unique values not found in the earlier data.

Table 5: Initial and last values for anonymised identifiers 2024

	Initial value	Last value	Number of new values
idno_anon	19360784	20362199	1,001,416
accessionno_anon	11734760	12596647	861,888
birthdate_anon	17740	18359	620
fname_anon	2754077	2872974	118,898
sname_anon	588699	616527	27,829
class_code	231472	537751	306,280

Details for class are included in Table 5 though class is a coded variable, and not a learner identifier. The coding of class would in theory permit the linking of classes over all the years, though this would be affected by whether a school or the system as a whole changed its class labelling system. There was clearly a change in 2024 relative to earlier years. The original 2024 labels were all new, and could not be found among the pre-2024 class labels. There was thus a change across the entire system, or in the data transfer processes, in the labelling of class. This explains why the number of new values for class_code is so high in Table 5.

5 Updated tables on the internal consistency of the data and identifiers

The following tables and graph follow the same methodologies as in the previous technical reports. The general picture that emerges is that marginal problems with respect to the utility of the identifiers remain noteworthy, and of a similar nature as in previous years. For instance, Gauteng's missing idno_anon problem continues to have the effect of compromising the linking of learners between Grade 7 in one year and in (above all) Grade 8 the next year – see Figure 1.

Table 6: Percentage of learners with unique identification per year

	A	B	C	D	E	A or D	A, B or D
idno	●						
accessionno		●					
birthdate			●	●	●		
fname			●	●			
sname			●	●	●		
gender			●	●	●		
race			●				
2017	84	35		85	62	85	90
2018	91	29	90	99	70	100	100
2019	92	37		99	70	100	100
2020	93	33	99	99	70	100	100
2021	92	36	99	99	69	99	100
2022	92	36	99	99	70	100	100
2023	92	36	99	99	69	100	100
2024	92	35	99	99	69	100	100

Table 7: Linking to next year

	A	B	D	A then D	A then D then B
All					
2017	67	14	61	75	76
2018	74	15	84	89	89
2019	84	26	85	90	90
2020	85	26	85	90	90
2021	84	29	86	90	90
2022	84	29	86	90	90
2023	84	28	85	90	90
Only grades 1 to 6					
2017	73	13	67	82	83
2018	80	15	92	96	96
2019	90	27	93	96	97
2020	90	27	93	96	97
2021	89	30	93	96	97
2022	89	30	94	96	97
2023	89	30	94	97	97

Minor changes in the 2021 and 2022 rows in the previous table, relative to the previous version of this report, are due to a correction undertaken when the current report was prepared.

In the next graph, the ‘A then D then B’ approach to linking has been used, with an additional condition requiring the grade movement to be only one grade up or the same grade. Any grade movement not complying with this condition is what would be described as ‘strange’ in Table 8. See the earlier report on the likelihood of some of these ‘strange’ differences being a true reflection of reality, as opposed to a data problem.

Figure 1: Degrees of linking by single grade and province for 2023

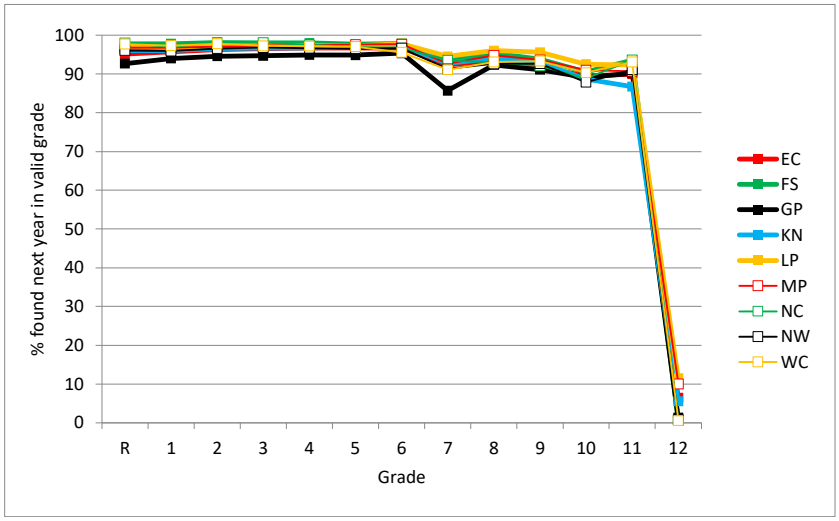

Table 8: Percentage of learners linked but to strange grade

	2017	2018	2019	2020	2021	2022	2023
EC	0.8	0.4	0.4	1.0	0.4	0.4	0.3
FS	0.4	0.3	0.3	0.4	0.5	0.4	0.3
GP	1.2	1.0	0.6	1.0	0.6	0.6	0.6
KN	1.3	1.1	0.7	0.8	0.6	1.1	0.8
LP	0.8	0.3	0.6	0.8	0.6	0.3	0.4
MP	2.1	0.4	0.5	1.0	0.9	0.5	0.4
NC	0.3	0.3	0.5	0.5	0.6	0.5	0.3
NW	0.3	0.4	0.3	0.6	0.7	0.3	0.4
WC	0.1	0.4	0.6	0.8	0.5	0.4	0.4
SA	1.0	0.6	0.5	0.8	0.6	0.6	0.5

Appendix 1: Stata code used for addition of 2024

```
* BRINGING TOGETHER PROVINCES INTO ONE NORMALISED FILE FOR 2024

* Before merging the nine provincial files I (a) removed a few lines of metadata at the bottom of the provincial files, (b) changed .rpt file extensions to .txt and (c) removed a few double quotation marks in several provincial files.
foreach p in "EC" "FS" "GP" "KZN" "LP" "MP" "NC" "NW" "WC" {
  display "`p' ********************************************************************************"
  import delimited "D:\Probably garbage 2\LURITS from LM 2025 04\\`p'_Learner_2024.txt", delimiter("#") encoding(UTF-8) clear
  keep emiscode learnerid accessionno sname fname birthdate idno gender grade class race
  local myvartype: type birthdate
  if substr("`myvartype'", 1, 3)=="str" {
    destring birthdate, force replace
  }
  local myvartype: type grade
  if substr("`myvartype'", 1, 3)=="str" {
    destring grade, force replace
  }
  local myvartype: type class
  if substr("`myvartype'", 1, 3)=="int" | substr("`myvartype'", 1, 3)=="lon" {
    tostring class, force replace
  }
  if "`p'"!="EC" {
    append using "C:\My Documents\Numbercrunching\LURITS\temp3.dta"
  }
  compress
  save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace
}
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
keep emiscode grade birthdate
tostring birthdate, gen(yob) // 178 missing
replace yob = substr(yob, 1, 4)
destring yob, replace
gen age = 2024 - yob
gen l = 1
destring grade, force replace
tostring emiscode, gen(statssaprov)
replace statssaprov = substr(statssaprov, 1, 1)
destring statssaprov, replace force // 6 missing - these 6 will have to be dropped, but that's all
merge m:1 statssaprov using "C:\My Documents\Numbercrunching\Tools\provinces.dta"
table grade myprov if grade>=0 & grade<=12, content(sum l) row col  // Looks plausible
table grade myprov, content(sum l) row col // Strange 101 to 114 grades spread fairly evenly across all provinces except WC.
codebook emiscode if grade>=101 & grade<=114 // 129 - too few to be special all schools, but check on EMIS number suggests this must be a sub-set of special schools.
table grade age if grade>=0 & grade<=12 & age>=5 & age<=15, content(sum l)
table grade age if grade>=101 & grade<=114 & age>=5 & age<=15, content(sum l) // This does not work like ordinary grades. I need to treat this differently, make 99.
table myprov  if grade>=0 & grade<=12, content(sum l) row col format(%15.0f) // 13626980 - BR annexure has 13501K for 2024. Difference of 125K is  close to what special school enrolment would be (some googling says we're currently around 140K). So it all tallies.
table myprov, content(sum l) row col format(%15.0f) // 13679374 - A further 50K in irregular grades.
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
format emiscode %9.0f
drop if emiscode==. // 6 - all have strange info
gen year = 2024
drop if emiscode<100000000 | emiscode>999999999 // 0
replace idno = subinstr(idno, " " , "", .) // 173
replace idno = upper(idno) // 290
replace idno = "" if idno=="NULL" // 114,350
replace idno = substr(idno, 1, 13) // 320
replace accessionno = subinstr(accessionno, " ", "", .) // 2,502
replace accessionno = upper(accessionno) // 65,695
foreach n of varlist fname sname {
  replace `n' = subinstr(`n', " ", "", .)
```

```
  replace `n' = substr(`n', 1, 20)
  replace `n' = upper(`n')
  rename `n' temp
  gen `n' = ""
  gen templength = length(temp)
  quietly summ templength
  quietly forvalues j = 1 / `r(max)' {
    replace `n' = `n' + substr(temp, `j', 1) if inrange(substr(temp, `j', 1), "A", "Z")
            noisily display `j'
  }
  replace `n' = substr(`n', 1, 15)
  drop temp*
}
replace grade = 99 if grade<0 | grade>12
codebook grade, tab(50) // 52,394 are 99
replace gender = upper(gender)
codebook gender, tab(500)
replace gender = cond(gender=="FEMALE", "F", cond(gender=="MALE", "M", ""))
replace race = upper(race)
replace race = subinstr(race, " ", "", .)
codebook race, tab(500)
rename race temp
gen race = "A" if strpos(temp, "AFRICAN")>0 | strpos(temp, "BLACK")>0
replace race = "C" if strpos(temp, "COLOURED")>0
replace race = "I" if strpos(temp, "INDIAN")>0 | strpos(temp, "ASIAN")>0
replace race = "W" if strpos(temp, "WHITE")>0
replace race = "O" if strpos(temp, "OTHER")>0
drop temp
codebook race, tab(50) // 1,228 missing
keep year emiscode idno accessionno birthdate fname sname grade class gender race
order year emiscode idno accessionno birthdate fname sname grade class gender race
compress
save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace

* MASTER LISTS

foreach p in "Eastern Cape" "Free State" "Gauteng" "KwaZulu Natal" "Limpopo" "Mpumalanga"
"Northern Cape" "North West" "Western Cape" "Special Needs  Education Centres" {
  if "`p'"=="Eastern Cape" {
    local pcode = "EC"
  }
  if "`p'"=="Free State" {
    local pcode = "FS"
  }
  if "`p'"=="Gauteng" {
    local pcode = "GT"
  }
  if "`p'"=="KwaZulu Natal" {
    local pcode = "KZN"
  }
  if "`p'"=="Limpopo" {
    local pcode = "LP"
  }
  if "`p'"=="Mpumalanga" {
    local pcode = "MP"
  }
  if "`p'"=="Northern Cape" {
    local pcode = "NC"
  }
  if "`p'"=="North West" {
    local pcode = "NW"
  }
  if "`p'"=="Western Cape" {
    local pcode = "WC"
  }
  if "`p'"=="Special Needs  Education Centres" {
    local pcode = "SNE"
```

```
  }
  display "Current is `p'
**********************************************************************************************************************
"
  import excel "C:\My Documents\Resources (Data)\Department of Education\Databases downloaded off
DoE website\Schools list for 2024 Q3 downloaded 2025 04\\`p'.xlsx", sheet("`pcode'") firstrow
case(lower) clear
  keep natemis sector type_doe
  if "`pcode'"!="EC" {
    append using "C:\My Documents\Numbercrunching\LURITS\temp4.dta"
  }
  save "C:\My Documents\Numbercrunching\LURITS\temp4.dta", replace
}
use "C:\My Documents\Numbercrunching\LURITS\temp4.dta", clear
rename natemis emiscode
format emiscode %9.0f
replace sector = upper(sector)
codebook sector
replace sector = substr(sector, 1, 1)
rename type_doe type
codebook type // 16 instances of "STAND ALONE ECD" will become "S"
replace type = substr(type, 1, 1)
keep emiscode sector type
rename sector Sector24
rename type Type24
egen tagging = tag(emiscode)
keep if tagging==1 // 0
drop tagging
codebook Type24 // N.B. no missing!!!
compress
save "C:\My Documents\Numbercrunching\LURITS\school2025_06.dta"

* THE AGGREGATES

use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
keep year emiscode grade
tostring emiscode, gen(statssaprov)
replace statssaprov = substr(statssaprov, 1, 1)
destring statssaprov, replace
merge m:1 statssaprov using "C:\My Documents\Numbercrunching\Tools\provinces.dta"
drop statssaprov
drop _merge
merge m:1 emiscode using "C:\My Documents\Numbercrunching\LURITS\school2025_06.dta"
drop if _merge==2
codebook emiscode if _merge==1 // 3 schools
drop _merge
gen ones = 1
count // 13,679,374 - official publication has 13,527,283 for ordinary. Difference of around 170K must be
special needs.
table myprov grade if grade>=0 & grade<=12 & Type24=="O", content(sum ones) format(%15.0f) row
col // Grade totals very close
table myprov grade if grade>=1 & grade<=12 & Type24=="O", content(sum ones) format(%15.0f) row
col // 12,649,413 - 0.2% lower than in official report, around 20K learners. Around 19K of the
discrepancy is in grades 8 to 12. And virtually all relates to GP!!!

* MISSING DATA

* By year...
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
quietly foreach v in "idno" "accessionno" "birthdate" "fname" "sname" "grade" "class" "gender" "race" {
  capture confirm variable `v'
  if _rc==0 {
            if "`v'"=="birthdate" | "`v'"=="grade" {
              if "`v'"=="birthdate" {
                        gen temp = cond(`v'==., 1, 0)
     }
              else {
```

```
            gen temp = cond(`v'==99, 1, 0)
          }
        }
        else {
          gen temp = cond(`v'=="", 1, 0)
        }
        mean temp
        local perc = _b[temp] * 100
        drop temp
        noisily display "`v' " _column(15) `perc'
  }
}
* By province...
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
tostring emiscode, replace
gen statssaprov = substr(emiscode, 1, 1)
destring statssaprov, replace
merge m:1 statssaprov using "C:\My Documents\Numbercrunching\Tools\provinces.dta"
drop _merge
quietly foreach v in "idno" "accessionno" "birthdate" "fname" "sname" "grade" "class" "gender" "race" {
  capture confirm variable `v'
  if _rc==0 {
        if "`v'"=="birthdate" | "`v'"=="grade" {
          if "`v'"=="birthdate" {
                gen missing = cond(`v'==., 1, 0)
          }
          else {
                gen missing = cond(`v'==99, 1, 0)
          }
        }
        else {
          gen missing = cond(`v'=="", 1, 0)
        }
  }
  noisily display "`v'"
  noisily tabstat missing, by(myprov)
  drop missing
}

* ANONYMISATION

* idno
use "C:\My Documents\Numbercrunching\LURITS\idno_key.dta", clear
append using "C:\My Documents\Numbercrunching\LURITS\idno_key 2022-2023.dta"
save "C:\My Documents\Numbercrunching\LURITS\temp4.dta", replace
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
gen temp = length(idno)
replace idno = "0" + idno if temp==12
replace idno = "00" + idno if temp==11
replace idno = "000" + idno if temp==10
drop temp
gen long idno_anon2 = .
order idno_anon2, after(idno)
merge m:1 idno using "C:\My Documents\Numbercrunching\LURITS\temp4.dta"
* <<<
summ idno_anon
local newstart = r(max) + 1
display "new start is " %15.0f `newstart' // 19360784
drop if _merge==2
replace idno_anon2 = idno_anon if _merge==3
drop _merge idno_anon
save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace
keep if idno!="" & idno_anon2==.
contract idno
drop _freq
gen myrand = uniform()
sort myrand
```

```
drop myrand
gen long idno_anon = `newstart' + _n - 1
summ idno_anon
local newend = r(max)
display "new end is "  %15.0f `newend' // 20362199 - an additional one million, seems plausible
* >>>
count // 1,001,416
save "C:\My Documents\Numbercrunching\LURITS\idno_key 2024.dta"
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
merge m:1 idno using "C:\My Documents\Numbercrunching\LURITS\idno_key 2024.dta"
count if _merge==3 & idno_anon2!=. // 0
replace idno_anon2 = idno_anon if _merge==3
drop idno idno_anon _merge
rename idno_anon2 idno_anon
compress
save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace
* accessionno
use "C:\My Documents\Numbercrunching\LURITS\accessionno_key.dta", clear
append using "C:\My Documents\Numbercrunching\LURITS\accessionno_key 2022-2023.dta"
save "C:\My Documents\Numbercrunching\LURITS\temp4.dta", replace
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
gen long accessionno_anon2 = .
order accessionno_anon2, after(accessionno)
merge m:1 accessionno using "C:\My Documents\Numbercrunching\LURITS\temp4.dta"
* <<<
summ accessionno_anon
local newstart = r(max) + 1
display "new start is " %15.0f `newstart' // 11734760
drop if _merge==2
replace accessionno_anon2 = accessionno_anon if _merge==3
drop _merge accessionno_anon
save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace
keep if accessionno!="" & accessionno_anon2==.
contract accessionno
drop _freq
gen myrand = uniform()
sort myrand
drop myrand
gen long accessionno_anon = `newstart' + _n - 1
summ accessionno_anon
local newend = r(max)
display "new end is " %15.0f `newend' //  12596647
* >>>
count // 861,888
save "C:\My Documents\Numbercrunching\LURITS\accessionno_key 2024.dta"
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
merge m:1 accessionno using "C:\My Documents\Numbercrunching\LURITS\accessionno_key
2024.dta"
count if _merge==3 & accessionno_anon2!=. // 0
replace accessionno_anon2 = accessionno_anon if _merge==3
drop accessionno accessionno_anon _merge
rename accessionno_anon2 accessionno_anon
compress
save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace
* birthdate (also insertion of birthyear)
use "C:\My Documents\Numbercrunching\LURITS\birthdate_key.dta", clear
append using "C:\My Documents\Numbercrunching\LURITS\birthdate_key 2022-2023.dta"
save "C:\My Documents\Numbercrunching\LURITS\temp4.dta", replace
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
tostring birthdate, gen(temp1)
gen tempyear = substr(temp1, 1, 4)
destring tempyear, replace force
gen tempmonth = substr(temp1, 5 , 2)
destring tempmonth, replace force
gen birthyear = tempyear + cond(tempmonth>6, .5, 0)
drop temp*
gen long birthdate_anon2 = .
```

```
order birthdate_anon2, after(birthdate)
merge m:1 birthdate using "C:\My Documents\Numbercrunching\LURITS\temp4.dta"
* <<<
summ birthdate_anon
local newstart = r(max) + 1
display "new start is " %15.0f `newstart' // 17740
drop if _merge==2
replace birthdate_anon2 = birthdate_anon if _merge==3
drop _merge birthdate_anon
save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace
keep if birthdate!=. & birthdate_anon2==.
contract birthdate
drop _freq
gen myrand = uniform()
sort myrand
drop myrand
gen long birthdate_anon = `newstart' + _n - 1
summ birthdate_anon
local newend = r(max)
display "new end is " %15.0f `newend' // 18359
* >>>
count // 620
save "C:\My Documents\Numbercrunching\LURITS\birthdate_key 2024.dta"
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
merge m:1 birthdate using "C:\My Documents\Numbercrunching\LURITS\birthdate_key 2024.dta"
count if _merge==3 & birthdate_anon2!=. // 0
replace birthdate_anon2 = birthdate_anon if _merge==3
drop birthdate birthdate_anon _merge
rename birthdate_anon2 birthdate_anon
compress
save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace
* fname and sname
foreach n in "fname" "sname" {
  use "C:\My Documents\Numbercrunching\LURITS\\`n'_key.dta", clear
  append using "C:\My Documents\Numbercrunching\LURITS\\`n'_key 2022-2023.dta"
  save "C:\My Documents\Numbercrunching\LURITS\temp4.dta", replace
  use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
  gen long `n'_anon2 = .
  order `n'_anon2, after(`n')
  merge m:1 `n' using "C:\My Documents\Numbercrunching\LURITS\temp4.dta"
  summ `n'_anon
  local newstart = r(max) + 1
  display "new start is " %15.0f `newstart' // 2754077, 588699
  drop if _merge==2
  replace `n'_anon2 = `n'_anon if _merge==3
  drop _merge `n'_anon
  save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace
  keep if `n'!="" & `n'_anon2==.
  contract `n'
  drop _freq
  gen myrand = uniform()
  sort myrand
  drop myrand
  gen long `n'_anon = `newstart' + _n - 1
  summ `n'_anon
  local newend = r(max)
  display "new end is " %15.0f `newend' // 2872974, 616527
  count // 118,898, 27,829
  save "C:\My Documents\Numbercrunching\LURITS\\`n'_key 2024.dta"
  use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
  merge m:1 `n' using "C:\My Documents\Numbercrunching\LURITS\\`n'_key 2024.dta"
  count if _merge==3 & `n'_anon2!=. // 0
  replace `n'_anon2 = `n'_anon if _merge==3
  drop `n' `n'_anon _merge
  rename `n'_anon2 `n'_anon
  compress
  save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace
```

```
}

* CODING CLASS, GENDER AND RACE

* class
use "C:\My Documents\Numbercrunching\LURITS\class_key.dta", clear
append using "C:\My Documents\Numbercrunching\LURITS\class_key 2022-2023.dta"
save "C:\My Documents\Numbercrunching\LURITS\temp4.dta", replace
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
gen long class_code2 = .
order class_code2, after(class)
merge m:1 class using "C:\My Documents\Numbercrunching\LURITS\temp4.dta"
* <<<
summ class_code
local newstart = r(max) + 1
display "new start is " %15.0f `newstart' // 231472
drop if _merge==2
replace class_code2 = class_code if _merge==3
drop _merge class_code
save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace
keep if class!="" & class_code==.
contract class
drop _freq
gen long class_code = `newstart' + _n - 1
summ class_code
local newend = r(max)
display "new end is " %15.0f `newend' // 537751
* >>>
count // 306,280
save "C:\My Documents\Numbercrunching\LURITS\class_key 2024.dta"
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
merge m:1 class using "C:\My Documents\Numbercrunching\LURITS\class_key 2024.dta"
count if _merge==3 & class_code2!=. // 0
replace class_code2 = class_code if _merge==3
drop class class_code _merge
rename class_code2 class_code
compress
save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace
* gender and race
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
rename gender temp
encode temp, gen(gender)
drop temp
rename race temp
encode temp, gen(race)
drop temp
codebook gender race // Looks right
order gender race, before(birthyear)
compress
save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace
* Now all done...
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
compress
save "C:\My Documents\Numbercrunching\LURITS\learner2025_06.dta"

* LEARNERS IDENTIFIED UNIQUELY WITHIN EACH YEAR

use "C:\My Documents\Numbercrunching\LURITS\learner2025_06.dta", clear
by year idno_anon, sort: egen temp = count(_n)
gen idno_id = cond(temp==1 & idno_anon!=., 1, 0)
drop temp
tabstat idno_id, by(year)
by year accessionno_anon, sort: egen temp = count(_n)
gen accessionno_id = cond(temp==1 & accessionno_anon!=., 1, 0)
drop temp
tabstat accessionno_id, by(year)
by year birthdate_anon fname_anon sname_anon gender race, sort: egen temp = count(_n)
```

```
gen combo1_id = cond(temp==1 & birthdate_anon!=. & fname_anon!=. & sname_anon!=. & gender!=. &
race!=., 1, 0)
drop temp
tabstat combo1_id, by(year)
by year birthdate_anon fname_anon sname_anon gender, sort: egen temp = count(_n)
gen combo2_id = cond(temp==1 & birthdate_anon!=. & fname_anon!=. & sname_anon!=. & gender!=.,
1, 0)
drop temp
tabstat combo2_id, by(year)
by year birthdate_anon sname_anon gender, sort: egen temp = count(_n)
gen combo3_id = cond(temp==1 & birthdate_anon!=. & sname_anon!=. & gender!=., 1, 0)
drop temp
tabstat combo3_id, by(year)
egen composite1 = rowmax(idno_id combo2_id)
tabstat composite1, by(year)
egen composite2 = rowmax(idno_id accessionno_id combo2_id)
tabstat composite2, by(year)

* LINKING TO NEXT YEAR

use "C:\My Documents\Numbercrunching\LURITS\learner2024_03.dta", clear
keep year emiscode idno_anon accessionno_anon birthdate_anon fname_anon sname_anon grade
gender race
keep if year==2023
append using "C:\My Documents\Numbercrunching\LURITS\learner2025_06.dta"
keep year emiscode idno_anon accessionno_anon birthdate_anon fname_anon sname_anon grade
gender race
gen L = 1
tabstat L if grade>=1 & grade<=6, stat(sum) by(year)
save "C:\My Documents\Numbercrunching\LURITS\temp11.dta", replace
* A: idno
use "C:\My Documents\Numbercrunching\LURITS\temp11.dta", clear
drop if idno_anon==.
by year idno_anon, sort: egen temp = count(_n)
keep if temp==1
drop temp
forvalues y = 2023 / 2023 {
  gen temp = 1 if year==`y' & grade>=1 & grade<=6
  by idno_anon, sort: egen prim`y' = max(temp)
  drop temp
}
keep idno_anon year prim* L
reshape wide L, i(idno_anon prim*) j(year)
quietly forvalues y = 2023 / 2023 {
  local yplus = `y' + 1
  gen templinked = 1 if L`y'==1 & L`yplus'==1
  summ templinked
  local all = r(sum)
  summ templinked if prim`y'==1
  local prim = r(sum)
  noisily display `y' _column(8) `all' _column(18) `prim'
  drop temp*
}
* B: accessionno
use "C:\My Documents\Numbercrunching\LURITS\temp11.dta", clear
drop if accessionno_anon==.
by year accessionno_anon, sort: egen temp = count(_n)
keep if temp==1
drop temp
forvalues y = 2023 / 2023 {
  gen temp = 1 if year==`y' & grade>=1 & grade<=6
  by accessionno_anon, sort: egen prim`y' = max(temp)
  drop temp
}
keep accessionno_anon year prim* L
reshape wide L, i(accessionno_anon prim*) j(year)
quietly forvalues y = 2023 / 2023 {
```

```
  local yplus = `y' + 1
  gen templinked = 1 if L`y'==1 & L`yplus'==1
  summ templinked
  local all = r(sum)
  summ templinked if prim`y'==1
  local prim = r(sum)
  noisily display `y' _column(8) `all' _column(18) `prim'
  drop temp*
}
* D: birthdate fname sname gender
use "C:\My Documents\Numbercrunching\LURITS\temp11.dta", clear
drop if birthdate_anon==. | fname_anon==. | sname_anon==. | gender==.
by year birthdate_anon fname_anon sname_anon gender, sort: egen temp = count(_n)
keep if temp==1
drop temp
forvalues y = 2023 / 2023 {
  gen temp = 1 if year==`y' & grade>=1 & grade<=6
  by birthdate_anon fname_anon sname_anon gender, sort: egen prim`y' = max(temp)
  drop temp
}
keep birthdate_anon fname_anon sname_anon gender year prim* L
reshape wide L, i(birthdate_anon fname_anon sname_anon gender prim*) j(year)
quietly forvalues y = 2023 / 2023 {
  local yplus = `y' + 1
  gen templinked = 1 if L`y'==1 & L`yplus'==1
  summ templinked
  local all = r(sum)
  summ templinked if prim`y'==1
  local prim = r(sum)
  noisily display `y' _column(8) `all' _column(18) `prim'
  drop temp*
}
* A then D
quietly forvalues y = 2023 / 2023 {
  local yplus = `y' + 1
  use "C:\My Documents\Numbercrunching\LURITS\temp11.dta", clear
  keep if year==`y' | year==`yplus'
  keep year grade idno_anon birthdate_anon fname_anon sname_anon gender
  gen long newid = _n
  save "C:\My Documents\Numbercrunching\LURITS\temp12.dta", replace
  keep year grade newid idno_anon
  drop if idno==.
  by year idno, sort: egen temp = count(_n)
  keep if temp==1
  drop temp
  by idno, sort: egen temp = count(_n)
  keep if temp==2
  drop temp
  save "C:\My Documents\Numbercrunching\LURITS\temp13.dta", replace
  use "C:\My Documents\Numbercrunching\LURITS\temp12.dta", clear
  merge 1:1 newid using "C:\My Documents\Numbercrunching\LURITS\temp13.dta", keepusing(newid)
  drop if _merge==3
  drop _merge
  keep year grade newid birthdate_anon fname_anon sname_anon gender
  drop if birthdate_anon==. | fname_anon==. | sname_anon==. | gender==.
  by year birthdate_anon fname_anon sname_anon gender, sort: egen temp = count(_n)
  keep if temp==1
  drop temp
  by birthdate_anon fname_anon sname_anon gender, sort: egen temp = count(_n)
  keep if temp==2
  drop temp
  count
  append using "C:\My Documents\Numbercrunching\LURITS\temp13.dta"
  count if year==`y'
  local all = r(N)
  count if year==`y' & grade>=1 & grade<=6
  local prim = r(N)
```

```
  noisily display `y' _column(8) `all' _column(18) `prim'
}
* A then D then B
quietly forvalues y = 2023 / 2023 {
  noisily display "Year is `y'
**************************************************************************************************"
  local yplus = `y' + 1
  use "C:\My Documents\Numbercrunching\LURITS\temp11.dta", clear
  keep if year==`y' | year==`yplus'
  tostring emiscode, replace force
  replace emiscode = substr(emiscode, 1, 1)
  destring emiscode, replace
  rename emiscode statssaprov
  merge m:1 statssaprov using "C:\My Documents\Numbercrunching\Tools\provinces.dta"
  noisily table grade myprov if grade!=99 & year==`y', content(sum L)
  keep year myprov grade idno_anon accessionno_anon birthdate_anon fname_anon sname_anon
gender
  gen long newid = _n
  save "C:\My Documents\Numbercrunching\LURITS\temp12.dta", replace
  use "C:\My Documents\Numbercrunching\LURITS\temp12.dta", clear
  keep year myprov grade newid idno_anon
  drop if idno==.
  by year idno, sort: egen temp = count(_n)
  keep if temp==1
  drop temp
  by idno, sort: egen temp = count(_n)
  keep if temp==2
  drop temp
  gen group = "A"
  sort idno year
  save "C:\My Documents\Numbercrunching\LURITS\temp13.dta", replace
  use "C:\My Documents\Numbercrunching\LURITS\temp12.dta", clear
  merge 1:1 newid using "C:\My Documents\Numbercrunching\LURITS\temp13.dta", keepusing(newid)
  drop if _merge==3
  drop _merge
  keep year myprov grade newid birthdate_anon fname_anon sname_anon gender
  drop if birthdate_anon==. | fname_anon==. | sname_anon==. | gender==.
  by year birthdate_anon fname_anon sname_anon gender, sort: egen temp = count(_n)
  keep if temp==1
  drop temp
  by birthdate_anon fname_anon sname_anon gender, sort: egen temp = count(_n)
  keep if temp==2
  drop temp
  gen group = "D"
  sort birthdate_anon fname_anon sname_anon gender year
  append using "C:\My Documents\Numbercrunching\LURITS\temp13.dta"
  save "C:\My Documents\Numbercrunching\LURITS\temp13.dta", replace
  use "C:\My Documents\Numbercrunching\LURITS\temp12.dta", clear
  merge 1:1 newid using "C:\My Documents\Numbercrunching\LURITS\temp13.dta", keepusing(newid)
  drop if _merge==3
  drop _merge
  keep year myprov grade newid accessionno_anon
  drop if accessionno_anon==.
  by year accessionno_anon, sort: egen temp = count(_n)
  keep if temp==1
  drop temp
  by accessionno_anon, sort: egen temp = count(_n)
  keep if temp==2
  drop temp
  gen group = "B"
  sort accessionno_anon year
  append using "C:\My Documents\Numbercrunching\LURITS\temp13.dta"
  gen gradeokay = 1 if year==`y' & (grade[_n + 1] - grade==1 | grade[_n + 1] - grade==0)
  count if year==`y'
  local all = r(N)
  count if year==`y' & grade>=1 & grade<=6
  local prim = r(N)
```

```
  count if year==`y' & gradeokay==1
  local gradechecked = r(N)
  noisily display `y' _column(8) `all' _column(18) `prim' _column(28) `gradechecked'
  save "C:\My Documents\Numbercrunching\LURITS\temp14.dta", replace
  keep if year==`y'
  drop if grade==99
  gen L = 1
  noisily table grade myprov, content(sum L)
  noisily table grade myprov if gradeokay==1, content(sum L)
}

* DIGGING DEEPER INTO THE LINKED WITH THE WRONG GRADE USING 2022-2023

use "C:\My Documents\Numbercrunching\LURITS\temp14.dta", clear
tabstat year if year==2022 & gradeokay!=1, by(group) stat(count) // 94207
keep if group=="A"
sort idno_anon year
by idno_anon, sort: egen temp = max(gradeokay)
keep if temp==.
drop temp
gen temp1 = grade if year==2022
by idno_anon, sort: egen tempstart = mean(temp1)
gen temp2 = grade if year==2023
by idno_anon, sort: egen tempend = mean(temp2)
gen gradediff = tempend - tempstart
drop temp*
replace gradediff = 85 if gradediff>85
replace gradediff = -85 if gradediff<-85
gen outofrange = cond(gradediff==85 | gradediff==-85, 1, 0)
tabstat outofrange if year==2022 // .2737428
codebook gradediff if year==2022 & outofrange!=1, tab(500)
use "C:\My Documents\Numbercrunching\LURITS\temp14.dta", clear
keep if group=="D"
sort birthdate_anon fname_anon sname_anon gender year
by birthdate_anon fname_anon sname_anon gender, sort: egen temp = max(gradeokay)
keep if temp==.
drop temp
gen temp1 = grade if year==2022
by birthdate_anon fname_anon sname_anon gender, sort: egen tempstart = mean(temp1)
gen temp2 = grade if year==2023
by birthdate_anon fname_anon sname_anon gender, sort: egen tempend = mean(temp2)
gen gradediff = tempend - tempstart
drop temp*
replace gradediff = 85 if gradediff>85
replace gradediff = -85 if gradediff<-85
gen outofrange = cond(gradediff==85 | gradediff==-85, 1, 0)
tabstat outofrange if year==2022 // .2454218
codebook gradediff if year==2022 & outofrange!=1, tab(500)
```