Addition of years 2022 and 2023 to the anonymised learner-level dataset

8 March 2024[1]

Contents

1 Introduction 1
2 The original source data 1
3 Merging and normalisation occurring before anonymisation 2
4 How 2022 and 2023 data were anonymised 3
5 Updated tables on the internal consistency of the data and identifiers 3
Appendix 1: Stata code used for addition of 2022 and 2023 6

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-2022-2023-v1	1,033,978	12	27,148,278
school-2022-2023-v1	202	5	25,142

The two tables are extensions of earlier tables finalised in December 2022, which contained details for the years 2017 to 2021. The extension is thus an extension of two years, and the abovementioned files contain data just for the years 2022 and 2023. The anonymisation process for 2022 and 2023 ensured anonymised identifiers for those two years were consistent with the ones from the earlier years. An analyst wanting to track learner movements from 2017 to 2023 would thus need to combine tables from the earlier work, in particular *learner2022_12.dta*, and the new *learner2024_03.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 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 current report should be read with that earlier report. The current report attempts not to repeat information from that earlier report.

2 The original source data

For the table *school2024_03*, new information was sourced from the following two master lists of schools:

2022	Title on DBE website: 'Quarter 3 of 2022: September 2022'
2023	Title on DBE website: 'Quarter 3 of 2023'

[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 2022 and 2023 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 very high. For instance, with respect to grades R to 12 in ordinary schools, public plus independent, in 2023 the total from the data of 13,415,754 is only 0.02% higher than the total from the official reports, the difference being 2,903 learners. For 2022, the difference is only 81 learners. The 'Other' in the data would be mostly special school learners, with the numbers roughly being in line with the corresponding values for 2017 to 2021.

Table 2: Official totals versus totals in the data

	2022			2023		
	Official	Data	%	Official	Data	%
EC	1,823,949	1,823,949	0.00	1,804,037	1,803,450	-0.03
FS	726,642	726,640	0.00	721,355	721,354	0.00
GP	2,601,899	2,601,832	0.00	2,617,592	2,617,686	0.00
KN	2,880,220	2,880,215	0.00	2,872,166	2,872,164	0.00
LP	1,796,729	1,796,728	0.00	1,797,818	1,800,648	0.16
MP	1,144,410	1,144,410	0.00	1,149,183	1,149,183	0.00
NC	305,336	305,336	0.00	305,758	306,330	0.19
NW	874,484	874,483	0.00	879,385	879,384	0.00
WC	1,241,238	1,241,233	0.00	1,265,557	1,265,555	0.00
Total	13,394,907	13,394,826	0.00	13,412,851	13,415,754	0.02
Other		166,963			170,735	
Final	13,419,971	13,561,789	1.06	13,439,683	13,586,489	1.09

For each of 2022 and 2023, there are no variables where the data are completely missing. The following two tables refer to the 11 variables of the *pre*-anonymised data, and where values are missing. On the whole, the completeness of the data has improved over time, though it is noteworthy that there were slightly more missing for *idno* in the last two years than in 2020 and 2021. This problem remains particularly large in Gauteng.

Table 3: Percentage missing by year

	2017	2018	2019	2020	2021	2022	2023
idno	13.3	8.5	7.5	6.6	6.8	7.3	7.3
accessionno	0.2	0.0	0.0	13.9	0.0	0.0	0.0
birthdate	13.4	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
sname	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
class	100.0	8.9	100.0	13.9	0.0	0.0	0.0
gender	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

Table 4: Percentage missing by province for 2022 and 2023 combined

	EC	FS	GP	KN	LP	MP	NC	NW	WC
idno	4.8	5.9	14.6	5.6	3.4	6.7	2.5	5.4	8.6
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	0.1	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	0.2	0.4	0.7	0.2	0.1	0.2	0.3	0.4	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	0.1	0.0

4 How 2022 and 2023 data were anonymised

The anonymisation process was relatively straightforward, and followed the methods employed for the 2017 to 2021 dataset. 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 2021 data was 17404361, meaning the new series of identifiers had to start at 17404362, and continued to 19360783, giving 1,956,422 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 2022 and 2023

	Initial value	Last value	Number of new values
idno_anon	17404362	19360783	1,956,422
accessionno_anon	9989763	11734759	1,744,997
birthdate_anon	16225	17739	1,515
fname_anon	2502998	2754076	251,079
sname_anon	528388	588698	60,311
class_code	131493	231471	99,979

Details for class are included in Table 5 though class is a coded variable, and not a learner identifier. The coding of class would permit the linking of classes over all the years, though this would be affected by whether a school changed its class labelling system.

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

The following tables and graphs follow the same methodologies as in the previous technical report. Details in this regard should be found there. 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 and Figure 2.

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

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	83	29	85	89	89
2022	83	28	85	89	89
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	90	31	95	97	98
2022	90	31	95	98	98

In the two graphs, 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 2021

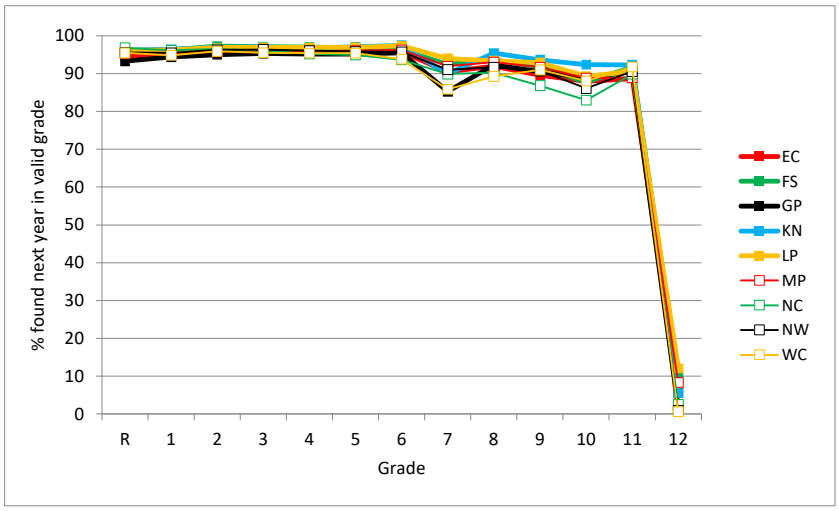

Figure 2: Degrees of linking by single grade and province for 2022

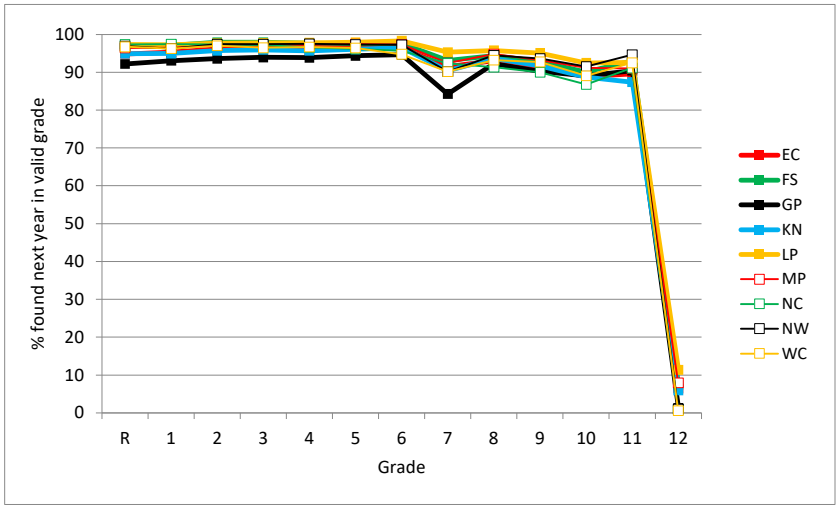

Table 8: Percentage of learners linked but to strange grade

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

Appendix 1: Stata code used for addition of 2022 and 2023

```
* BRINGING TOGETHER ONE NORMALISED FILE COVERING 2022 AND 2023

* 2022
import delimited "C:\My Documents\Resources (Data)\Department of Education\LURITS 2022 and 2023
obtained 2024 02\Learners_2022.txt", delimiter("#") varnames(1) case(lower) encoding(UTF-8) clear
* I checked, and the last learner here is the same as the last year when I view in WordPad.
destring emiscode, replace force
format emiscode %9.0f
drop if emiscode==. // 6 - 2 of these were admin rows at the very end of the original
count // 13,561,789 - bottom line in official publication is 13,419,971
gen year = 2022
drop if emiscode<100000000 | emiscode>999999999 // 0
replace idno = subinstr(idno, " " , "", .) // 651
replace idno = upper(idno) // 294
replace idno = "" if idno=="NULL" // 3,887
replace idno = substr(idno, 1, 13) // 278
replace accessionno = subinstr(accessionno, " ", "", .) // 1,590
replace accessionno = upper(accessionno) // 61,786
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*
}
destring birthdate, replace force
codebook grade, tab(100)
destring grade, force replace
replace grade = 99 if grade<0 | grade>12
codebook grade, tab(50) // 43,689 are 99
replace gender = upper(gender)
codebook gender, tab(50)
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 substr(temp, 1, 6)=="AFRICA" | substr(temp, 1, 5)=="BLACK"
replace race = "C" if substr(temp, 1, 8)=="COLOURED"
replace race = "I" if substr(temp, 1, 6)=="INDIAN" | substr(temp, 1, 5)=="ASIAN"
replace race = "W" if substr(temp, 1, 5)=="WHITE"
replace race = "O" if substr(temp, 1, 5)=="OTHER"
drop temp
codebook race, tab(50) // 2,336 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
* 2023
import delimited "C:\My Documents\Resources (Data)\Department of Education\LURITS 2022 and 2023
obtained 2024 02\Learners_2023.txt", delimiter("#") varnames(1) case(lower) encoding(UTF-8) clear
* Couldn't compare to WordPad as the latter wouldn't open, but it seems clear enough I have all the
data. .
destring emiscode, replace force
format emiscode %9.0f
drop if emiscode==. // 5 - 2 of these were admin rows at the very end of the original. I checked and the
other 3 were not 'concertina' observations with lots of data.
```

```
count // 13,586,489 - bottom line in official publication is 13,439,683
gen year = 2023
drop if emiscode<100000000 | emiscode>999999999 // 0
replace idno = subinstr(idno, " " , "", .) // 445
replace idno = upper(idno) // 257
replace idno = "" if idno=="NULL" // 110,372!!!
replace idno = substr(idno, 1, 13) // 238
replace accessionno = subinstr(accessionno, " ", "", .) // 2,237
replace accessionno = upper(accessionno) // 64,121
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*
}
destring birthdate, replace force
codebook grade, tab(100)
destring grade, force replace
replace grade = 99 if grade<0 | grade>12
codebook grade, tab(50) // 48,304 are 99
replace gender = upper(gender)
codebook gender, tab(50)
replace gender = "F" if gender=="FEMALE"
replace gender = "M" if gender=="MALE"
replace gender = "" if gender!="F" & gender!="M" // 53
replace race = upper(race)
replace race = subinstr(race, " ", "", .)
codebook race, tab(500)
rename race temp
gen race = "A" if substr(temp, 1, 6)=="AFRICA" | substr(temp, 1, 5)=="BLACK"
replace race = "C" if substr(temp, 1, 8)=="COLOURED"
replace race = "I" if substr(temp, 1, 6)=="INDIAN" | substr(temp, 1, 5)=="ASIAN"
replace race = "W" if substr(temp, 1, 5)=="WHITE"
replace race = "O" if substr(temp, 1, 5)=="OTHER"
drop temp
codebook race, tab(50) // 1,173 missing
tostring class, replace
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
append using "C:\My Documents\Numbercrunching\LURITS\temp3.dta"
save "C:\My Documents\Numbercrunching\LURITS\temp3.dta", replace

* MASTER LISTS

use "C:\My Documents\Resources (Data)\Department of Education\Databases downloaded off DoE
website\Schools list for 2022 Q3 downloaded 2023 07\National (1).dta", clear
rename natemis emiscode
destring emiscode, gen(temp) force
keep if temp>=100000000 & temp<=999999999 // 6
rename sector Sector
replace Sector = upper(Sector)
codebook Sector
replace Sector = substr(Sector, 1, 1)
rename type_doe Type
codebook Type
replace Type = substr(Type, 1, 1)
keep emiscode Sector Type
```

```
rename Sector Sector22
rename Type Type22
destring emiscode, replace
format emiscode %9.0f
egen tagging = tag(emiscode)
keep if tagging==1 // 1
drop tagging
codebook Type22 // N.B. no missing!!!
save "C:\My Documents\Numbercrunching\LURITS\temp0.dta", replace
import excel "C:\My Documents\Resources (Data)\Department of Education\Databases downloaded off
DoE website\Schools list for 2023 Q3 downloaded 2024 03\National - ordinary schools.xlsx",
sheet("Ordinary School") firstrow clear
keep if NatEmis>=100000000 & NatEmis<=999999999 // 0
format NatEmis %9.0f
rename NatEmis emiscode
replace Sector = upper(Sector)
codebook Sector
replace Sector = substr(Sector, 1, 1)
rename Type_DoE Type
codebook Type
replace Type = substr(Type, 1, 1)
keep emiscode Sector Type
rename Sector Sector23
rename Type Type23
codebook Type23 // N.B. no missing!!!
merge 1:1 emiscode using "C:\My Documents\Numbercrunching\LURITS\temp0.dta"
drop _merge
order emiscode Sector22 Type22 Sector23 Type23
compress
save "C:\My Documents\Numbercrunching\LURITS\school2024_03.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\school 2024_03.dta"
drop if _merge==2
drop _merge
gen ones = 1
count if year==2022
table myprov if year==2022 & Type22=="O" & grade>=0 & grade<=12, content(sum ones)
count if year==2023
table myprov if year==2023 & Type23=="O" & grade>=0 & grade<=12, content(sum ones)

* MISSING DATA

* By year...
forvalues y = 2022 / 2023 {
  use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
  keep if year==`y'
  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\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\idno_key.dta"
* <<<
summ idno_anon
local newstart = r(max) + 1
display "new start is " %15.0f `newstart' // 17404362
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' // 19360783 - an additional two million, seems plausible
```

```
* >>>
count // 1,956,422
save "C:\My Documents\Numbercrunching\LURITS\idno_key 2022-2023.dta"
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
merge m:1 idno using "C:\My Documents\Numbercrunching\LURITS\idno_key 2022-2023.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\temp3.dta", clear
gen long accessionno_anon2 = .
order accessionno_anon2, after(accessionno)
merge m:1 accessionno using "C:\My Documents\Numbercrunching\LURITS\accessionno_key.dta"
* <<<
summ accessionno_anon
local newstart = r(max) + 1
display "new start is " %15.0f `newstart' // 9989763
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' // 11734759
* >>>
count // 1,744,997
save "C:\My Documents\Numbercrunching\LURITS\accessionno_key 2022-2023.dta"
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
merge m:1 accessionno using "C:\My Documents\Numbercrunching\LURITS\accessionno_key 2022-
2023.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\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\birthdate_key.dta"
* <<<
summ birthdate_anon
local newstart = r(max) + 1
display "new start is " %15.0f `newstart' // 16225
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' // 17739
* >>>
count // 1,515
save "C:\My Documents\Numbercrunching\LURITS\birthdate_key 2022-2023.dta"
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
merge m:1 birthdate using "C:\My Documents\Numbercrunching\LURITS\birthdate_key 2022-2023.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\temp3.dta", clear
  gen long `n'_anon2 = .
  order `n'_anon2, after(`n')
  merge m:1 `n' using "C:\My Documents\Numbercrunching\LURITS\\`n'_key.dta"
  * <<<
  summ `n'_anon
  local newstart = r(max) + 1
  display "new start is " %15.0f `newstart' // 2502998
  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' // 11734759
  * >>>
  count // 1,744,997
  save "C:\My Documents\Numbercrunching\LURITS\\`n'_key 2022-2023.dta"
  use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
  merge m:1 `n' using "C:\My Documents\Numbercrunching\LURITS\\`n'_key 2022-2023.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\temp3.dta", clear
gen long class_code2 = .
order class_code2, after(class)
merge m:1 class using "C:\My Documents\Numbercrunching\LURITS\class_key.dta"
* <<<
summ class_code
local newstart = r(max) + 1
```

```
display "new start is " %15.0f `newstart' // 131493
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' // 11734759
* >>>
count // 1,744,997
save "C:\My Documents\Numbercrunching\LURITS\class_key 2022-2023.dta"
use "C:\My Documents\Numbercrunching\LURITS\temp3.dta", clear
merge m:1 class using "C:\My Documents\Numbercrunching\LURITS\class_key 2022-2023.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 // Same as in earlier data
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\learner2024_03.dta"

* LEARNERS IDENTIFIED UNIQUELY WITHIN EACH YEAR

use "C:\My Documents\Numbercrunching\LURITS\learner2024_03.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\learner2022_12.dta", clear
keep year emiscode idno_anon accessionno_anon birthdate_anon fname_anon sname_anon grade
gender race
keep if year==2021
append using "C:\My Documents\Numbercrunching\LURITS\learner2024_03.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 = 2021 / 2022 {
  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 = 2021 / 2022 {
  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 = 2021 / 2022 {
  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 = 2021 / 2022 {
  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 = 2021 / 2022 {
  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 = 2021 / 2022 {
  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 = 2021 / 2022 {
  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 = 2021 / 2022 {
  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)
```