**MARYLAND APPROVED PROGRAM COMPLETER SYSTEM**

**(MAPCS)**

**MHEC Internal Validation Code**

**STATA Syntax Code**

**Posted 10/03/2019**

# Data Validation Syntax

The following Stata code is used to validate all file submissions. Additional validation testing may be completed contingent upon the results of this baseline code. New validation steps may be added over time as the scope of the collection changes.

\*file population\*

\*compare totals for each tab to prior years\*

\*compare sic to fice for correct sic construction\*

table fice /\*also known as OPEID\*?

table sic /\*MHEC constructed institution level identifier\*/

table degree

table mdpgcode

\*count unique populations\*

\*first instance of each unit record within an fice is marked as unique, additional instances as duplciate\*

\*add to excel tab and cross check against prior year totals\*

sort fice idn

egen num\_unique=tag(fice idn)

label define counts 1 "unique" 0 "duplicate"

label values num\_unique counts

tab num\_unique

tab fice if num\_unique==1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*validate FICE-SIC alignment\*

gen sicfice=.

replace sicfice=1 if fice==143600& sic==260700

replace sicfice=1 if fice==205700& sic==110100

replace sicfice=1 if fice==205800& sic==110200

replace sicfice=1 if fice==206100& sic==111300

replace sicfice=1 if fice==206200& sic==120600

replace sicfice=1 if fice==206300& sic==111250

replace sicfice=1 if fice==206400& sic==111000

replace sicfice=1 if fice==206500& sic==263300

replace sicfice=1 if fice==206700& sic==261200

replace sicfice=1 if fice==206800& sic==121400

replace sicfice=1 if fice==207100& sic==111700

replace sicfice=1 if fice==207200& sic==121800

replace sicfice=1 if fice==207300& sic==262000

replace sicfice=1 if fice==207400& sic==112100

replace sicfice=1 if fice==207500& sic==112200

replace sicfice=1 if fice==207600& sic==262300

replace sicfice=1 if fice==207700& sic==262500

replace sicfice=1 if fice==207800& sic==262600

replace sicfice=1 if fice==208000& sic==262700

replace sicfice=1 if fice==208300& sic==143000

replace sicfice=1 if fice==208600& sic==263100

replace sicfice=1 if fice==208900& sic==113600

replace sicfice=1 if fice==209100& sic==123900

replace sicfice=1 if fice==209200& sic==263700

replace sicfice=1 if fice==209500& sic==154000

replace sicfice=1 if fice==209900& sic==124200

replace sicfice=1 if fice==210200& sic==124400

replace sicfice=1 if fice==210300& sic==124700

replace sicfice=1 if fice==210400& sic==124500

replace sicfice=1 if fice==210500& sic==124600

replace sicfice=1 if fice==210600& sic==124800

replace sicfice=1 if fice==210700& sic==265000

replace sicfice=1 if fice==210800& sic==265200

replace sicfice=1 if fice==210900& sic==265400

replace sicfice=1 if fice==465000& sic==111100

replace sicfice=1 if fice==691100& sic==112970

replace sicfice=1 if fice==817500& sic==112400

replace sicfice=1 if fice==830800& sic==110900

replace sicfice=1 if fice==1001400& sic==111900

replace sicfice=1 if fice==1164400& sic==124900

replace sicfice=1 if fice==2073900& sic==115470

replace sicfice=1 if fice==3100700& sic==110770

tab sicfice

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*Step 2: Duplicate Check\*

\*this must be done first as variables will be constructed throughout validation that will

\*alter results when looking for duplicates across all variables\*

sort fice idn

\*dups on all variables\*

\*no records should be identical on all elements\*

duplicates tag, generate(dups)

tab dups

\*validate counts are same on idn, tssn and tssn2 to make sure idn is correctly converted\*

\*dups on ssn only\*

duplicates tag idn, generate (dupssn)

duplicates tag tssn, generate (duptssn)

duplicates tag tssn2, generate (duptssn2)

tab dupssn

tab duptssn

tab duptssn2

\*duplicate checking matching on ssn, fice, mapcs code\*

\*no records should be included twice with same code\*

sort fice idn mdpgcode

duplicates tag idn fice mdpgcode, generate(ssn\_pg)

duplicates tag tssn fice mdpgcode, generate(tssn\_pg)

duplicates tag tssn2 fice mdpgcode, generate(tssn2\_pg)

tab ssn\_pg

tab tssn\_pg

tab tssn2\_pg

tab fice ssn\_pg if ssn\_pg!=0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*validate that when multiple records are present, key data elements are identical in all records\*

\*static elements shifting, must encode string hispanic\*

\*other elements may change between mapcs award (entry year, entry term, georg, res, etc. are specific to the MAPCS code, demographics should not change\*

encode hispanic, gen (hislat)

tab hispanic

tab hislat

bysort fice idn: generate chcit = citizen - citizen[\_n-1]

bysort fice idn: generate chhis = hislat - hislat[\_n-1]

bysort fice idn: generate chwh = rwhite - rwhite[\_n-1]

bysort fice idn: generate chbk = rblack - rblack[\_n-1]

bysort fice idn: generate chas = rasian - rasian[\_n-1]

bysort fice idn: generate chai = raian - raian[\_n-1]

bysort fice idn: generate chnh = rnhpi- rnhpi[\_n-1]

bysort fice idn: generate chsex = sex - sex[\_n-1]

bysort fice idn: generate chidt = idt - idt[\_n-1]

bysort fice idn: generate chflg = mapcsflg - mapcsflg[\_n-1]

tab chcit /\*0 is no change, all other numbers are measure of change\*/

tab fice if (chcit!=0& chcit!=.) & num\_unique==0 /\*changer\*/

tab chhis /\*0 is no change, all other numbers are measure of change\*/

tab fice if (chhis!=0&chhis!=.) & num\_unique==0 /\*changer\*/

tab chwh /\*0 is no change, all other numbers are measure of change\*/

tab fice if (chwh!=0 &chwh!=.)& num\_unique==0 /\*changer\*/

tab chbk /\*0 is no change, all other numbers are measure of change\*/

tab fice if (chbk!=0&chbk!=.)& num\_unique==0 /\*changer\*/

tab chas /\*0 is no change, all other numbers are measure of change\*/

tab fice if (chas!=0 & chas!=.) & num\_unique==0 /\*changer\*/

tab chai /\*0 is no change, all other numbers are measure of change\*/

tab fice if (chai!=0 & chai!=.) & num\_unique==0 /\*changer\*/

tab chnh /\*0 is no change, all other numbers are measure of change\*/

tab fice if (chnh!=0 &chnh!=.) & num\_unique==0 /\*changer\*/

tab chsex /\*0 is no change, all other numbers are measure of change\*/

tab fice if (chsex!=0&chsex!=.) & num\_unique==0 /\*changer\*/

tab chidt /\*0 is no change, all other numbers are measure of change\*/

tab fice if (chidt!=0&chidt!=.) & num\_unique==0 /\*changer\*/

tab chflg /\*0 is no change, all other numbers are measure of change\*/

tab fice if (chflg!=0&chflg!=.) & num\_unique==0 /\*changer\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*validate individual data elements\*

\*valid fice, 1 valid, -9 not valid\*

gen vfice=.

replace vfice=1 if fice==143600 | fice==205700 | fice==205800 | fice==206100 | fice==206200 | fice==206300 | fice==206400 | ///

fice==206500 | fice==206700 | fice==206800 | fice==207100 | ///

fice==207200 | fice==207300 | fice==207400 | fice==207500 | fice==207600 | fice==207700 | ///

fice==207800 | fice==208000 | fice==208600 | fice==208700 | fice==208900 | ///

fice==209100 | fice==209200 | fice==209500 | fice==209600 | fice==209900 | fice==210200 | fice==210300 | ///

fice==210400 | fice==210500 | fice==210600 | fice==210700 | fice==210800 | fice==210900 | fice==465000 | ///

fice== 691100 | fice==795900 | fice==817500 | fice==830800 | fice==1001400 | fice==1164400 | fice== 2073900 | ///

fice==208300 | fice==2127900 | fice==2314800 | fice==2578400 | fice==3100700 | fice==3455500 | fice==90795900 | fice==99765900

mvdecode vfice, mv(-9)

tab vfice /\*valid fice, 1 valid, -9 not valid\*/

tab fice if vfice==-9 /\*list of invalid fice numbers\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*valid sic, 1 valid, -9 not valid\*

gen vsic=.

replace vsic=1 if sic==110100 | sic==110200 | sic==110770 | sic==110900 | sic==111000 | ///

sic==111100| sic==111250| sic==111300 | sic==111700 | sic==111900 | sic==112100 | ///

sic==112200 | sic==112400| sic==112970 | sic==113600 | sic==115470| sic==120600 | ///

sic==121400 | sic==121800| sic==123900| sic==124200 | sic==124400| sic==124500 | ///

sic==124600 | sic==124700 | sic==124800| sic==124900 | sic==124960 | sic==124965 | ///

sic==143000 | sic==154000 | sic==260400 | sic==260410 | sic==260469 | sic==260537| ///

sic==260700 | sic==262000 | sic==262050 | sic==262230 | sic==262300| sic==262487| ///

sic==262500| sic==262570 | sic==262600| sic==262650 | sic==262700 | sic==262767 | ///

sic==263100 | sic==263124 | sic==263200| sic==263300| sic==263500 | sic==263700| ///

sic==263915 | sic==264100| sic==264150 | sic==264182| sic==264187 | sic==264300 | ///

sic==265000| sic==261200| sic==265100 | sic==265200 | sic==265300| sic==265400| ///

sic==265420| sic==265600 | sic==261575| sic==263575| sic==260500 | sic==261150| ///

sic==261450| sic==110800| sic==111500 | sic==111600 | sic==112800| sic==112900| ///

sic==112950 | sic==260250

mvdecode vsic, mv(-9)

tab vsic, missing /\*1 valid, -9 not valid\*/

tab sic if vsic==-9 /\*list of invalid sic numbers\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*valid collection term, 1 is valid, 2 is not valid\*

gen vct = .

\*select one of the following\*

replace vct = 1 if coltm == 8

replace vct = 2 if coltm !=8

tab vct /\*1 is valid, 2 is not valid\*/

tab fice if vct==2 /\*1 is valid, 2 is not valid\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*valid collection year, 1 is valid, 2 is not valid\*

\*change year as needed\*

gen vcy = .

\*select one of the following\*

replace vcy = 1 if colyr == 2018

replace vcy = 2 if colyr != 2018

replace vcy = 1 if colyr == 2019

replace vcy = 2 if colyr != 2019

replace vcy = 1 if colyr == 2020

replace vcy = 2 if colyr != 2020

tab vcy /\*1 is valid, 2 is not valid\*/

tab fice if vcy==2 /\*1 is valid, 2 is not valid\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*sex, 1 valid, 2 not valid\*

gen vsex=.

replace vsex=1 if sex==1| sex==2| sex==3| sex==4

replace vsex=2 if sex>=5 | sex==. | sex<1

tab sex

tab vsex /\*1 is valid, 2 is not valid\*/

tab fice if vsex==2 /\*1 is valid, 2 is not valid\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*citizenship, 1 valid, 2 not valid\*

gen vcit=.

replace vcit=1 if citizen==1|citizen==2

replace vcit=2 if citizen<1 | citizen>2 | citizen==.

tab citizen

tab vcit /\*1 is valid, 2 is not valid\*/

tab fice if vcit==2 /\*1 is valid, 2 is not valid\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*race/ethnicity, 1 valid, 2 not valid\*

gen vhisp=.

replace vhisp=1 if hislat==1 | hislat==2 |hislat==.

replace vhisp=2 if hislat!=1 & hislat!=2 & hislat!=.

mvencode hislat, mv(9)

tab hispanic

tab hislat

tab vhisp /\*1 is valid, 2 is not valid\*/

tab fice if vhisp==2 /\*1 is valid, 2 is not valid\*/

gen vrwhite=.

replace vrwhite=1 if rwhite==1 |rwhite==.

replace vrwhite=2 if rwhite!=1 & rwhite!=.

tab vrwhite /\*1 is valid, 2 is not valid\*/

tab fice if vrwhite==2 /\*1 is valid, 2 is not valid\*/

gen vrblack=.

replace vrblack=1 if rblack==2|rblack==.

replace vrblack=2 if rblack!=2 & rblack!=.

tab vrblack /\*1 is valid, 2 is not valid\*/

tab fice if vrblack==2 /\*1 is valid, 2 is not valid\*/

gen vrasian=.

replace vrasian=1 if rasian==3 |rasian==.

replace vrasian=2 if rasian!=3 & rasian!=.

tab vrasian /\*1 is valid, 2 is not valid\*/

tab fice if vrasian==2 /\*1 is valid, 2 is not valid\*/

gen vamer=.

replace vamer=1 if raian==4 |raian==.

replace vamer=2 if raian!=4 & raian!=.

tab vamer /\*1 is valid, 2 is not valid\*/

tab fice if vamer==2 /\*1 is valid, 2 is not valid\*/

gen vhawi=.

replace vhawi=1 if rnhpi==5 |rnhpi==.

replace vhawi=2 if rnhpi!=5 & rnhpi!=.

tab vhawi /\*1 is valid, 2 is not valid\*/

tab fice if vhawi==2 /\*1 is valid, 2 is not valid\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*birthdt - must be 8 digits with at least a valid 4 digit year + 0000 or can be 00000000 if unknown\*

\*considering birthdt between 1900 and 2017 valid, tagging unknowns as 3 and 4\*

gen vbirth=.

replace vbirth=1 if birthdt >=19170000 & birthdt <=20100000

replace vbirth=2 if birthdt <19170000 & birthdt!=0

replace vbirth=2 if birthdt >20100000 & birthdt!=0

replace vbirth=3 if birthdt==.

replace vbirth=4 if birthdt==0

tab vbirth /\*1 is valid, 2 is out of range, 3 is blank, 4 is 0\*/

tab fice vbirth if vbirth>1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

gen vmilstat=.

replace vmilstat=1 if milstat>=1 & milstat <=5

replace vmilstat=2 if milstat==.

replace vmilstat=3 if milstat>5

replace vmilstat=4 if milstat<1

tab milstat

tab vmilstat /\*1 is valid, 2 is blank, 3 is out of range\*/

tab fice if vmilstat>1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*geographic orgin\*

\*001 - foreign, 100 - unknown US state, 101-199, 200- maryland unknown count, 201-224, 000- unknown\*

gen vgeorg=.

replace vgeorg=1 if georg==1

replace vgeorg=3 if georg==100

replace vgeorg=4 if georg==0

replace vgeorg=5 if georg==200

replace vgeorg=6 if georg>200 & georg<=224

replace vgeorg=7 if georg>100 & georg<=199

replace vgeorg=2 if georg==.|georg>224|georg<0

tab vgeorg

tab fice if vgeorg==3 /\*100 - unknown US state\*/

tab fice if vgeorg==4 /\*unknown\*/

tab fice if vgeorg==5 /\*Unknown MD county\*/

tab fice if vgeorg==2 /\*value out of range\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*mdresidency code\*

gen vmdres=.

replace vmdres=1 if mdres==1|mdres==2

replace vmdres=3 if mdres>2

replace vmdres=2 if mdres==.

replace vmdres=3 if mdres<1

tab vmdres /\*1 is valid, 2 is blank, 3 is out of range\*/

tab fice if vmdres>1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*valid ID Type, valid 1, invalid 2\*

gen vtype=.

replace vtype = 1 if idt ==1 | idt==2| idt==3

replace vtype = 2 if idt !=1 & idt !=2 & idt!=3 & idt !=.

tab vtype /\*1 is valid, 2 not valid\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*valid SSN, 1 is valid, 2 is 8 digits, 3 is 10 or more digits, 4 is 7 or less digits\*

\*8 digits may indicate missing leading 0\*

gen vssn =idn

tostring vssn, replace

gen vssn2=.

replace vssn2=1 if strlen(vssn)==9

replace vssn2=2 if strlen(vssn)==8

replace vssn2=3 if strlen(vssn)>=10

replace vssn2=4 if strlen(vssn)<=7

tab vssn2

\*validate against id type\*

tab vssn2 if idt==1 &(vssn2==3|vssn2==4) /\*number of invalid IDNs with ID Type =SSN\*/

tab fice if idt==1 &(vssn2==3|vssn2==4) /\*fice with number of invalid IDNs with ID Type =SSN\*/

tab vssn2 idt

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*validate TSSN\*

\*tssn should be 10 characters\*

gen vtssn=.

replace vtssn=3 if strlen(tssn)>10

replace vtssn=2 if strlen(tssn)<10

replace vtssn=1 if strlen(tssn)==10

tab vtssn /\*1 is valid, 2 & 3 not valid\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*validate TSSN2\*

\*tssn2 should be 7 characters\*

gen vtssn2=.

replace vtssn2=3 if strlen(tssn2)>7

replace vtssn2=2 if strlen(tssn2)<7

replace vtssn2=1 if strlen(tssn2)==7

tab vtssn2 /\*1 is valid, 2 & 3 not valid\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*valid campusID 1 valid, 2 not valid\*

\*campus id is 10 or less characters\*

gen vcamp=.

replace vcamp=2 if strlen(tssn)>=11

replace vcamp=1 if strlen(tssn)<=10

tab vcamp /\*1 is valid, 2 not valid\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*valid SASID, code missing values as 9\*

gen vsasid =sasid

mvencode vsasid, mv(9)

tostring vsasid, replace

\*valid SASID\*, 1 and 2 are valid, 3 and 4 are not valid

gen vsasid2=.

replace vsasid2=4 if strlen(vsasid)>10 /\* not valid, too long\*/

replace vsasid2=3 if strlen(vsasid)<10 &strlen(vsasid)>1 /\* not valid, too short\*/

replace vsasid2=2 if strlen(vsasid)==10 /\*valid\*/

replace vsasid2=1 if strlen(vsasid)==1 /\*valid coded as 9 missing value\*/

tab vsasid2

tab fice if vsasid2>=3 /\*fice with SASID not valid\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*Valid degree sought, 1 valid, 2 not valid\*

gen vdegree=.

replace vdegree=1 if degree==10 | degree==20 | degree==30 | degree==40 | degree==50 | degree==60| ///

degree== 65|degree== 70| degree==81| degree==85| degree==86| degree==99 |degree==47 | degree==87 | degree==0

replace vdegree=2 if degree ==.

replace vdegree=3 if degree !=10 & degree!=20 & degree!=30 & degree!=40 & degree!=50 & ///

degree!=60 & degree!= 65 & degree!= 70 & degree!=81 & degree!=85 & degree!=86 & degree!=99 & degree!=47 & degree!=87 & degree!=0

tab degree vdegree

tab fice if vdegree>1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*valid Degree Date\*

\*update sequence to aling to expected graduating year for collection\*

\*collection is 9/1 to 8/31\*

\*9/1/17 to 8/31/18 for MAPCS18\*

gen vddate=.

replace vddate=1 if degdate !=. & (degdate>=20170901 & degdate<=20180831) /\*valid\*/

replace vddate=4 if degdate ==. /\*blank invalid\*/

replace vddate=2 if degdate <=20170831 /\*date too early for collection range\*/

replace vddate=3 if degdate >=20180901 /\*date too late for collection range\*/

tab vddate

tab fice vddate if vddate>1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*entry term\*

gen veterm=.

replace veterm=1 if entrm==1| entrm==2| entrm==3| entrm==4 /\*valid\*/

replace veterm=2 if entrm!=1 & entrm!=2& entrm!=3& entrm!=4 & entrm!=. /\*invalid\*/

replace veterm=3 if entrm==. /\*blank invalid\*/

tab veterm

tab fice if veterm>1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*entry year\*

gen veyear=.

replace veyear=1 if (entyr >=1990 & entyr <2018) /\*acceptable range\*/

replace veyear=2 if entyr<=1989 /\*invalid\*/

replace veye=2 if entyr>=2018 /\*invalid\*/

replace veyear=3 if entyr==. /\*blank invalid\*/

tab veyear

tab fice veyear if veyear>=2

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*validate MAPCS code is from current list\*

\*old codes may be in file if student is completing discontinued program\*

gen vmdpgcode=.

replace vmdpgcode= 1 if mdpgcode==517

replace vmdpgcode= 1 if mdpgcode==300

replace vmdpgcode= 1 if mdpgcode==1561

replace vmdpgcode= 1 if mdpgcode==1580

replace vmdpgcode= 1 if mdpgcode==1303

replace vmdpgcode= 1 if mdpgcode==412

replace vmdpgcode= 1 if mdpgcode==1213

replace vmdpgcode= 1 if mdpgcode==1305

replace vmdpgcode= 1 if mdpgcode==1214

replace vmdpgcode= 1 if mdpgcode==1581

replace vmdpgcode= 1 if mdpgcode==1250

replace vmdpgcode= 1 if mdpgcode==335

replace vmdpgcode= 1 if mdpgcode==104

replace vmdpgcode= 1 if mdpgcode==1215

replace vmdpgcode= 1 if mdpgcode==1230

replace vmdpgcode= 1 if mdpgcode==105

replace vmdpgcode= 1 if mdpgcode==1242

replace vmdpgcode= 1 if mdpgcode==1560

replace vmdpgcode= 1 if mdpgcode==249

replace vmdpgcode= 1 if mdpgcode==1312

replace vmdpgcode= 1 if mdpgcode==1200

replace vmdpgcode= 1 if mdpgcode==1231

replace vmdpgcode= 1 if mdpgcode==1203

replace vmdpgcode= 1 if mdpgcode==1557

replace vmdpgcode= 1 if mdpgcode==1568

replace vmdpgcode= 1 if mdpgcode==1311

replace vmdpgcode= 1 if mdpgcode==1232

replace vmdpgcode= 1 if mdpgcode==1582

replace vmdpgcode= 1 if mdpgcode==1583

replace vmdpgcode= 1 if mdpgcode==1584

replace vmdpgcode= 1 if mdpgcode==421

replace vmdpgcode= 1 if mdpgcode==1309

replace vmdpgcode= 1 if mdpgcode==1534

replace vmdpgcode= 1 if mdpgcode==1244

replace vmdpgcode= 1 if mdpgcode==1331

replace vmdpgcode= 1 if mdpgcode==1332

replace vmdpgcode= 1 if mdpgcode==1338

replace vmdpgcode= 1 if mdpgcode==1339

replace vmdpgcode= 1 if mdpgcode==1316

replace vmdpgcode= 1 if mdpgcode==1319

replace vmdpgcode= 1 if mdpgcode==1222

replace vmdpgcode= 1 if mdpgcode==1223

replace vmdpgcode= 1 if mdpgcode==1237

replace vmdpgcode= 1 if mdpgcode==1210

replace vmdpgcode= 1 if mdpgcode==324

replace vmdpgcode= 1 if mdpgcode==1233

replace vmdpgcode= 1 if mdpgcode==321

replace vmdpgcode= 1 if mdpgcode==1207

replace vmdpgcode= 1 if mdpgcode==428

replace vmdpgcode= 1 if mdpgcode==527

replace vmdpgcode= 1 if mdpgcode==1234

replace vmdpgcode= 1 if mdpgcode==1236

replace vmdpgcode= 1 if mdpgcode==1205

replace vmdpgcode= 1 if mdpgcode==340

replace vmdpgcode= 1 if mdpgcode==343

replace vmdpgcode= 1 if mdpgcode==342

replace vmdpgcode= 1 if mdpgcode==341

replace vmdpgcode= 1 if mdpgcode==385

replace vmdpgcode= 1 if mdpgcode==387

replace vmdpgcode= 1 if mdpgcode==1323

replace vmdpgcode= 1 if mdpgcode==420

replace vmdpgcode= 1 if mdpgcode==1734

replace vmdpgcode= 1 if mdpgcode==1313

replace vmdpgcode= 1 if mdpgcode==1241

replace vmdpgcode= 1 if mdpgcode==1586

replace vmdpgcode= 1 if mdpgcode==9999

replace vmdpgcode=2 if mdpgcode==. /\*blank, not valid\*/

tab vmdpgcode

tab fice vmdpgcode if vmdpgcode==2 /\*blank, not valid\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*population distributions\*\*\*\*\*\*\*\*\*\*\*\*\*

tab idt /\*high levels of 2 requires review\*/

tab citizen

tab sex

\*race/ethnicity\*

tab hispanic, missing

\*construct multi race\*

gen rmulti=.

replace rmulti=1 if (rwhite==1 & rblack==2 & rasian==3 & raian==4 & rnhpi==5 & hislat!=2)

replace rmulti=1 if (rwhite==1 & rblack==2 & hislat!=2)

replace rmulti=1 if (rwhite==1 & rasian==3 & hislat!=2)

replace rmulti=1 if (rwhite==1 & raian==4 & hislat!=2)

replace rmulti=1 if (rwhite==1 & rnhpi==5 & hislat!=2)

replace rmulti=1 if (rblack==2 & rasian==3 & hislat!=2)

replace rmulti=1 if (rblack==2 & raian==4 & hislat!=2)

replace rmulti=1 if (rblack==2 & rnhpi==5 & hislat!=2)

replace rmulti=1 if (rasian==3 & raian==4 & hislat!=2)

replace rmulti=1 if (rasian==3 & rnhpi==5 & hislat!=2)

replace rmulti=1 if (raian==4 & rnhpi==5 & hislat!=2)

tab rmulti

\*unknown race\*

gen runknown=.

replace runknown=2 if rwhite==. & rblack==. & rasian==. & raian==. & rnhpi==. & hislat!=2

tab runknown /\*high levels of unknown requires review\*/

\*\*distribution of population by race\*\*

tab rwhite if hislat!=2 & rmulti!=1

tab rblack if hislat!=2 & rmulti!=1

tab hispanic

tab hislat

tab rasian if hislat!=2 & rmulti!=1

tab raian if hislat!=2 & rmulti!=1

tab rnhpi if hislat!=2 & rmulti!=1

tab rmulti

tab runknown

tab degree

table georg if georg==001

table georg if georg==100

tabstat georg if georg>100 & georg <200, statistics( count )

table georg if georg==200

tabstat georg if georg>200 & georg <225, statistics( count )

table georg if georg==000

tab mdres

tab milstat

tab mapcsflg /\*high rate of 9999 requires review\*/

tabulate mdpgcode, nolabel

\*population issues\*

tab fice if idt!=1 /\*id provided is not SSN\*/

tab fice if idt!=1 & citizen==1 /\*id provided is not SSN & student is US citizen\*/

tab fice if mdpgcode==9999 /\*no mapcs code\*/

tab fice if runknown==2 /\*race/ethnicity unknown\*/

tab fice if vgeorg==3|vgeorg==4 /\*100 - unknown US state\*//\*unknown\*/

tab fice if vgeorg==5 /\*Unknown MD county\*/

\*prep DIS\*

drop coltm colyr subcamp sasid ptax fmjrd cumgpa cncrhre ccrhra pdflg entrm

drop entyr revtrn rqcrhr teachcan sex citizen hispanic-tssn\_pg dup2-num\_unique2

drop sic idt

rename program\_tax\_orig program

\*parse program\*

format %06.0f program

tostring program, gen (program2)

\*construct majord and disp\*

\*first two digits are MAJORDS\*

\*first four digits are DISCPLN\*

\*MAJORDS, first two digits\*

gen majords=substr(program2, 1, length(program2)-4)

destring majords, replace

\*DISCPLN - first four digits\*

gen discpln=substr(program2, 1, length(program2)-2)

destring discpln, replace

gen prog=discpln

\*drop program2\*

drop program2

\*step 1: convert to string\*

tostring majorlk, gen(fmjrd)

\*start with the first number and keep 2 digits\*

\*this pulls out the degree sought code from the first\_major string\*

gen fmdeg=substr(fmjrd, 1, 2)

\*start with the third digit and keep that digit and 3 additional)

\*this pulls out the program taxonomy code from the first\_major string\*

gen fmpt=substr(fmjrd,3, 4)

\*finally, destring to reinstate numbers for validation both degree code and program taxonomy\*

destring fmdeg fmpt, replace

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*tag potential teachers\*

\*these are likely to be in the MAPCS file\*

gen teach=.

replace teach=1 if fmpt==801|fmpt==802|fmpt==803|fmpt==804

replace teach=1 if fmpt==805|fmpt==806|fmpt==807|fmpt==808

replace teach=1 if fmpt==809|fmpt==810|fmpt==811

replace teach=1 if fmpt==812|fmpt==813|fmpt==814|fmpt==815

replace teach=1 if fmpt==816|fmpt==817|fmpt==818|fmpt==819

replace teach=1 if fmpt==820|fmpt==823

replace teach=1 if fmpt==830|fmpt==831|fmpt==832

replace teach=1 if fmpt==833|fmpt==834|fmpt==835|fmpt==836

replace teach=1 if fmpt==837|fmpt==838|fmpt==839

replace teach=1 if prog==801|prog==802|prog==803|prog==804

replace teach=1 if prog==805|prog==806|prog==807|prog==808

replace teach=1 if prog==809|prog==810|prog==811

replace teach=1 if prog==812|prog==813|prog==814|prog==815

replace teach=1 if prog==816|prog==817|prog==818|prog==819

replace teach=1 if prog==820|prog==823

replace teach=1 if prog==830|prog==831|prog==832

replace teach=1 if prog==833|prog==834|prog==835|prog==836

replace teach=1 if prog==837|prog==838|prog==839

tab fice teach if program!=.

\*restrict population to in scope degree date\*

\*update dates each year\*

gen inscope=.

replace inscope=1 if teach!=. & (degdate>=20170901 & degdate<=20180831)

tab inscope

tab fice if inscope==1

tab teach inscope

tab fice inscope if teach==1

\*save and drop all observations that are not teach=1 and in-scope=1\*

drop if teach!=1

drop if inscope!=1

\*identify unique possible teachers in DIS\*

gsort fice idn inscope

egen teach\_unique=tag(fice idn)

tab teach\_unique

\*keep only one record per idn/fice\*

drop if teach\_unique!=1

drop tssn campusid degdate degree program majorlk fmjrd majords discpln fmdeg fmpt prog teach inscope teach\_unique

\*this is final population of potential unit records for MAPCS\*

tab fice

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*prep MAPCS\*

drop subcamp sasid sex citizen hispanic rwhite rblack rasian raian rnhpi entrm

drop entyr birthdt georg mdres milstat sector segment num\_unique sicfice-runknown

drop coltm colyr tssn tssn2 idt campusid degdate degree sic source\_file

\*next drop duplicate records in MAPCS for testing MAPCS flag\*

\*mapcs\*

sort fice idn

egen map\_unique=tag(fice idn)

tab map\_unique

tab fice if map\_unique==1

drop if map\_unique==0

drop map\_unique

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*merge files for analysis\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

use "M:\AnnKellogg\Data Validation Analysis\_Ann Kellogg\Core Collections\DIS-MAPCS Merged Data Files\ MAPCS2018\_ready for merge2.dta saved"

merge 1:1 fice idn using "M:\AnnKellogg\Data Validation Analysis\_Ann Kellogg\Core Collections\DIS-MAPCS Merged Data Files\DIS2018\_ready for merge2.dta"

\*save data set\*

\*tab number of students in DIS NOT in MAPCS\*

\*these are potential missing records from MAPCS as these\*

\*students graduated with a HEGIS code and degree date\*

\*that make them potential MAPCS candidates\*

tab fice if \_merge==2

\*records in MAPCS with flag set to DIS=Yes

tab fice if mapcsflg==1

\*record in MAPCS with DIS flag=Yes but no matching\*

\*record found in DIS. Either flag is wrong, records is missing from DIS\*

\*or IDNs are not consistent across DIS and MAPCS to support merge\*

tab fice mapcsflg if \_merge==1 & mapcsflg==1

\*tab number of students with MAPCS Flag=Yes that found a match in DIS\*

tab fice mapcsflg if \_merge==3 & mapcsflg==1

\*number in mapcs\*

tab fice if mapcsflg!=.

\*tab number of students in DIS with MAPCS Flag=No\*

\*if any records are present then mapcsflg is incorrectly set\*

tab fice mapcsflg if \_merge==3 & mapcsflg==2