

Affordability Study Data Cleaning and Computed Variables

August 28, 2008

Data cleaning

1. As files were submitted by the institutions, variable frequencies, means and maximums were generated and sent back to institutions for review to verify the correct labeling of variables and that maximum values that were within program limits. Appropriate revisions of several institutions' data submissions were made.
2. Files were then matched against type 1 enrollment submissions for fall 2005-summer 2006, and cases that did not appear as undergraduate students in type 1 were flagged. Reviewing the matches, if the unmatched records represented less than 4 percent of the institution's undergraduate enrollment for the period, the unmatched records were dropped without consulting the institution. If the number was more than 4 percent, the institution was notified and a resolution was made on a case-by-case basis. The number and percent unmatched for each institution and any resolution that was negotiated with the institution is given in the table below:

Institution	Number unmatched	Percent UG enroll unmatched	Resolution if over 4% unmatched
Public Insts.			
Eastern	66	0.3%	
KSU	14	0.5%	
Morehead	3	<0.1%	
Murray	3	<0.1%	
UK	225	0.8%	
U of L	1	<0.1%	
NKU	113	1%	
WKU	6	<0.1%	
KCTCS	1,238	1%	
AIKCU Insts.			
Bellermine	569	21.5%	All but 139 were graduate students, all were dropped
Campbellsville	424	21.6%	Were graduate students -- dropped
Centre	4		
Georgetown	11	0.5%	
Ky. Wesleyan	18	2.3%	
Lindsey Wilson	806	32%	Were graduate students -- dropped
MidContinent	5	0.4%	
Midway	193	13%	Missing spring enrollments were added, reduced to 26 unmatched students who were dropped.
Pikeville	2	0.2%	
Thomas More	157	9.7%	Were graduate students -- dropped
Union	24	3.4%	
U. Cumberlands	36	2%	
St. Catherine	24	3%	

3. After merging the cleaned affordability data with existing enrollment data, students with problematic undergraduate status were removed to ensure that aid levels reflect “standard” undergraduate students only. 495 students who moved to graduate status during the year were removed, as were 21,176 high school students taking college courses (dual enrollment students). An additional 378 students at public institutions were dropped because they changed residency status at some time during the year, making their cost data difficult to interpret.
4. Other caveats and data oddities are as follows: 1) Union College does not track KY tuition grants separately in their system but reported this data as KHEAA work study; 2) Northern reported their institutional work study separately from the need/merit schema (variable = NKUwork); 3) Murray flagged concurrent enrollment students with “*” and a single Katrina student with “#” (variable = Musuast).
5. The cleaned CPE dataset was merged with KHEAA’s data for the same students, matching on SSN. For the affordability study, this data is only used to fill in missing data, most notably the income and other FAFSA fields of students in the independent sector.
6. The resulting final dataset includes all undergraduate students from the institutions listed above for whom enrollment was reported to the Council’s comprehensive database between fall 2005 and summer 2006, regardless of aid status, without the exclusions listed above.

Cases by Sector	Frequency	Percent
Public Research Universities	37,707	17%
Public Comprehensive Universities	68,442	31%
Public Two-Year Colleges	98,421	44%
AICKU Independents	17,866	8%
Total	222,436	100%

Computed Variables

Aid Variables:

```
StateNeedGrant = CAP + KTG + KYTEACH;
StateNonNeedGrant = KEES + ERLYCHLD + KYNATGRD + KYMNRTYED;
StateGrant = StateNeedGrant + StateNonNeedGrant;

StateAid = stategrant + kheaawork;

FedNeedGrant = PELL + FEDSEOGGR + FEDHLTHDGR + biagr;
FedNonNeedGrant = BYRDGR + rotcgr + jobvoc + veterans;
FedGrant = FedNeedGrant + FedNonNeedGrant;

FedNonNeedLoan = unsubstaff + plusln;
FedNeedLoan = substaff + perkinsln + fedhlthln;
FedLoan = FedNonNeedLoan + FedNeedLoan;

FedAid = FedGrant + FedLoan + Fedwork;

InstGrant = instndgr + instndmrgr + instmrgr + tuitwaiv;
InstLoan = instndln + instndmrln + instmrln;
InstWork = instndwrk + instndmrwrk + instmrwrk + nkuwork;

InstAid = INSTGRANT + INSTLOAN + INSTWORK;

OtherGrant = thrdprtygr + emptuit + othstgr;

OtherAid = othergrant + otherln + kapt;

TotalAid = stateaid + fedaid + instaid + otheraid;

TotalGrant = stategrant + fedgrant + instgrant + othergrant;
TotalLoan = fedloan + instloan + otherln;
TotalWork = kheaawork + fedwork + instwork;
TotalOther = kapt;

*other categories for detailed aid report;
otherfedgr = FEDHLTHDGR + biagr + rotcgr + jobvoc;
otherstgr = ERLYCHLD + KYNATGRD + KYMNRTYED + KYTEACH + othstgr;
```

Student Characteristics Variables [(BYR (birth year) – ZIP (zip code))]

These variables are student characteristics that are constants such as birth year, gender and race. They are collected for all enrolled students each semester of enrollment on the Type 1 enrollment file, defined in the CDB guidelines. Although these student characteristics are generally unchanging, missing data or data collection and reporting error may result in different values being reported to the council for the same student in different semesters. The following business rules were developed to reconcile differing values on these variables.

1. If a student was enrolled in the fall 2005 semester, the value submitted in fall 2005 was used.
2. If a student was not enrolled in the fall 2005 semester but was enrolled in the spring 2006 semester, then the spring 2006 value was used.
3. If a student was not enrolled in either the fall 2005 or spring 2006 semester, but was enrolled in the summer 2006 semester, then the summer 2006 value was used.

4. The variable “t1” gives the semester file from which these values were pulled.

Student Enrollment Variables (Resf05 – enrolr06)

These variables may change for each semester of a student’s enrollment, so the dataset includes this information at the semester level. The last three characters of the variable name give the semester: fall 2006 = f06, spring 2006 = s06, and summer 2006 = r06.

Retention and degrees conferred

Students who received a degree or other credential in 2005-06 or who were enrolled in fall of 2006 are counted as retained. Students were retained at the institutional level if they were enrolled in any way at the same semester the following fall. Students were retained at the system level if they enrolled in any reporting institution in Kentucky the following fall. The degrees referenced in the degree fields are the highest degree the student attained during the 2006-07 academic year.

Attendance Status

This variable was coded using the same logic that the National Postsecondary Student Aid Study used to code attendance status, but replaced the number of months enrolled with the number of semesters enrolled. If a student was full time for two or three semesters during the academic year, they are coded as full-time, full year. If a student was enrolled full-time for one semester and was not part-time any other semesters, the student was coded as full-time, part-year. Students enrolled full-time one semester and part-time for two or more semesters and students enrolled part-time for two or more semesters were coded as part-time, full-year. The remaining students were coded part-time, part-year. Because the aid amounts are for the full academic year and not particular semesters, most aid analysis considers only the first category above (full-time, full-year) to be full-time students, and lumps the remaining students into a part-time/part-year category.

Student Feedback File Variables (CYEAR – GPA1yr)

Variables 154-192 are from the student feedback file and include information on entrance and placement exam scores, GPA and hours earned at the end of the first year, and grades earned in the first developmental and college-level math and English courses. This information is collected for only first-time, degree or credential-seeking students at public institutions who entered in the summer or fall of 2005. This data was cleaned for discrepant values and matched against the first-time degree-seeking cohorts from each institution on the type 1 enrollment records for summer and fall of 2005. Only KCTCS had significant amounts of missing cases on the student feedback file when compared with type 1.

The following variables summarize whether or not a student is categorized as having developmental needs in a given subject, based on Kentucky’s mandatory placement policy which states that students must receive developmental education or further placement testing if they received a 17 or lower on the ACT or an equivalent score on the SAT or another standardized placement exam such as Compass or Accuplacer. These variables combine any test scores the student received into a single variable:

1. UNPRMATH – developmental needs in math; 1 = has developmental needs in math, 0 = does not have developmental needs in math
2. UNPRENG – developmental needs in English; 1 = has developmental needs in English, 0 = does not have developmental needs in English
3. UNPRREAD – developmental needs in reading; 1 = has developmental needs in reading, 0 = does not have developmental needs in reading
4. NUMBUNPR – number of developmental needs; 0 - 3 based on preceding three variables
5. UNDERPREP – whether or not a student is underprepared in one or more subjects, 1 = yes, 0 = no

Cost of attendance

Institutions reported the cost of attendance (CostAtt) at the student-unit level for students with aid. However, students without aid also need a cost of attendance to calculate price variables, and institutions use widely varying estimates of some elements of the total cost of attendance. Therefore, a modified version of the cost of attendance reported to IPEDS by the institutions for the 2005-06 academic year was used in determining cost and price.

The COA for in and out-of-state students were calculated by adding the full-time tuition and fees, books, and room and board reported by institutions to IPEDS for the 2005-06 AY, and then adding a standard amount for other or indirect expenses. If different room and board costs were reported to IPEDS for on and off-campus students, the average of these two amounts was used. The standard amount for other or indirect expenses is based on the institutions' urban or non-urban status. Other expenses for students at urban institutions match the other expense plus one-half of the transportation expense from the College Board's low 9-month living expense budget for regional cities in 2005-06 (Cincinnati, Cleveland and St. Louis). One-half of the transportation expense was included to estimate the average amount that on and off-campus students would pay. Non-urban institutions' other expenses are the weighted average of the on and off-campus amounts reported to IPEDS in 2005-06 by institutions by sector (weighted by fall 2005 undergraduate FTE for publics and fall 2005 headcount enrollment for independents). The resulting amounts for "other" expenses are \$3,469 for students at urban public or independent sector institutions, \$1,869 for students at non-urban public institutions, and \$2,075 for students at non-urban independent institutions.

For part-time students, costs were based on the per-credit-hour tuition multiplied by the number of hours a student attempted during the year (not hours completed), and one-third of the full academic year room, board and other costs. This pro-rating of one-third for non-tuition costs was based on the mean number of hours attempted by part-time students in the study during the academic year, which was 10.1 or approximately one-third of a standard 30-hour full-time schedule.

"COAFinal" is the variable with these amounts. This variable is used in most of the output produced by CPE. The exception is for fine-grained examinations of net price, stacking or other analyses for which this artificial COA will produce false over-awards (like the NASH analysis of unmet need). Use the actual cost of attendance reported by the institution or KHEAA, CostAtt, for these analyses. The third COA variable in the dataset, COAIPEDS, is simply the IPEDS cost amounts for full-time students with a calculation for part-time students. This last variable was never used because we wanted a number that was more comparable across institutions and so developed COAFinal.

Price Variables

The price variables were coded as follows, using the cost of attendance reported by the institutions or the IPEDS amount as calculated above if not reported by the institution. KAPT, although rarely reported, was included in the “total other” variable, and was applied to the costs in the first variable, netprice.

```
netprice = coafinal - totalgrant - totalother;
```

```
outofpocket = netprice - totalloan;  
if totalloan = . then outofpocket = netprice;
```

```
outofpockwork = outofpocket - totalwork;  
if totalwork = . then outofpockwork = outofpocket;
```

```
NetwoLoans = netprice - totalwork;
```

```
if efc < 3851 then pellelig = 1;  
else pellelig = 0;  
if efc = . then pellelig = .;
```

DirectCost = Tuition, fees and books as reported to IPEDS in 2005-06, using one-third of the book cost for part-time students

Income Quartiles

The income quartile variable “instqrtrle” includes two separate state quartiles, dependent students ranked by total adjusted family income and independent students ranked by total adjusted family income. The quartiles divide each of these populations into four equal categories. A fifth category is for students who did not fill out the FAFSA and for whom we have no family income data. An EFC quartile was also calculated, although the large number of ties of EFC = 0 places more than 25% of the relevant students into the first EFC quartile.

The incomes associated with the income quartiles are:

	Dependent Students	Independent Students
Lowest income quartile	Less than \$29,742	Less than \$7,326
Second income quartile	\$29,742 to \$51,982	\$7,326 to \$15,874
Third income quartile	\$51,983 to \$81,364	\$15,874 to \$29,223
Highest income quartile	\$81,365 and above	\$29,223 and above

Income Quintiles

The income quintile variables are done differently – they are based on state income quartiles from the 2006 ACS, not on the income distribution of enrolled students. The tables on which these are based are here: [KY HH inc quintiles, 2006 ACS.xlsx](#)

Unmet need

These variables are coded as follows:

$\text{Unmetneedwloans} = \text{CostAtt} - \text{EFC} - \text{totalgrants} - \text{totalother}$

$\text{Unmetneedwloans} = \text{CostAtt} - \text{EFC} - \text{totalgrants} - \text{totalother} - \text{totalloans}$

These variables use the institutionally-reported student-level COA because they were developed for an institutional key indicator, and this measure of cost is the best for use in comparing institutions' financial aid policies.

Categorical aid variables

This set of variables is coded 1 = yes and 0 = no whether or not a student received aid of that type. These variable are used to report the percent of students in a category who received aid of a given type.

When to use zeros

The deved06.final dataset has zeros in the aid fields if students did not receive aid of that type. When running summary stats like mean or median it is important to think through whether to keep these zeros or replace them with blanks/missing. Zeros are needed to do math with these fields (like adding and subtracting to create unmet need), because doing math with a missing value will produce a missing value in SAS. However, means and medians will include students with no aid if the zeros are kept, which is not usually what is intended. If the average grant aid of recipients is needed (what is usually meant by average grant aid) then the zeros need to be replaced by missing/blank.

There are three fields in the dataset for which zeros are valid values and should never be replaced with missing/blank. EFC can legitimately be 0 and often is. Also, family and student income (incfam and incstu) can legitimately be zero and even negative. Missing values in these fields usually means that the student did not file the FAFSA, although some is just missing even for FAFSA filers.