TEXAS HIGHER EDUCATION OPPORTUNITY PROJECT

Administrative College Transcript Data

Documentation for Public Use Data Files

November 21, 2008

Texas Higher Education Opportunity Project / Wallace Hall / Office of Population Research / Princeton University /
Princeton, NJ 08544 / Telephone: (609) 258-8185 / Fax: (609) 258-1039 / Email: theop@opr.princeton.edu

CONTENTS

OVERVIEW 1
Introduction 1
Confidentiality 1
Variable Availability and Comparability 2

VARIABLE DESCRIPTIONS 4

MERGING COLLEGE TRANSCRIPT WITH APPLICATION DATA 9

APPENDIX A – MERGING INSTRUCTIONS 10

APPENDIX B – FIELDS OF MAJOR AND DEPARTMENTS OF MAJOR BY FIELD 11

OVERVIEW

INTRODUCTION

The Texas Higher Education Opportunity Project (THEOP) Administrative College Transcript Public Use Data (henceforth College Transcript Data) is available for applicants who were admitted and subsequently enrolled at each of the nine THEOP universities. Each college transcript registers academic progress toward a degree for *a single enrollee in a single semester*, and specifically, provides hours earned, semester GPA, cumulative GPA, and department and field of major.

Because most enrollees attend college for more than one semester, there are usually multiple college transcripts for each enrollee. Table 1 below shows years and sizes for Application and College Transcript Data. For example, at UT Austin, 87,156 freshman applicants enrolled between 1991 and 2003. Between 1991 and 2005, this same group of enrollees generated 659,102 college transcripts. Note that for most institutions, College Transcript Data extends beyond the corresponding Application Data by one or more years.

Table 1. Years and Sizes of Application Data and College Transcript Data

	College Application Data			College Transcript Data	
University	Years Available	Number of Applications	Number of Enrollees	Years Available	Number of College Transcripts
Texas A&M	1992-2002	163,027	70,580	1992-2007	637,028
Texas A&M Kingsville	1992-2002	18,872	13,241	1992-2004	91,106
UT Arlington	1994-2002	29,844	12,041	1994-2002	51,315
UT Austin	1991-2003	210,006	87,156	1991-2004	659,102
UT Pan American	1995-2002	44,747	19,057	1995-2005	115,812
UT San Antonio	1990-2004	61,221	34,245	1990-2004	160,604
Texas Tech	1995-2003	81,153	30,573	1995-2004	211,771
Rice	2000-2004	36,190	3,430	2000-2005	18,149
Southern Methodist	1998-2005	45,549	11,422	1998-2005	60,607

It is not possible to track transfers within this group of nine universities. Identification codes (studentids) are unique to each university and the same studentid in two different college transcript files does not represent the same enrollee. Therefore, if building a dataset that combines College Transcript Data from more than one institution, creation of an institution identification code (institutionid) variable is recommended.

CONFIDENTIALITY

Our goal in preparing the College Transcript Data is to make available the highest quality information while minimizing the risk of applicants being identified. Several confidentiality measures are taken to achieve this goal, including:

- *Elimination of identifiers*: University-assigned applicant identification numbers are changed to new, randomly generated applicant identification numbers. Under special circumstances, researchers may apply for access to restricted use data through the THEOP website (http://theop.princeton.edu).
- *Collapsing small frequency cells*: Small frequency cells are eliminated by collapsing multiple values into range categories. For example, semester grade point averages below 0.6 are collapsed into the category 0 – 0.5.
- *Setting values to missing*: Some values are set to missing to preserve the well-established categories but to hide individual values. This strategy is used infrequently to minimize its impacts on analyses.

VARIABLE AVAILABILITY AND COMPARABILITY

Almost every variable in the College Transcript Data is available for all nine institutions, as shown in Table 2. The only exception is the variable GPA Hours (gpahrs) which is available only for Texas A&M.

Table 2. College Transcript Data: Variable Availability

Variable Name	Variable Label	Institution / Year								
		Texas A&M 1992-2007	Texas A&M Kingsville 1992-2004	UT Arlington 1994-2002	UT Austin 1991-2004	UT Pan American 1995-2005	UT San Antonio 1990-2004	Texas Tech 1995-2004	Rice 2000-2005	Southern Methodist 1998-2005
studentid	Student id	am	amk	ar	au	pa	sa	tt	ri	smu
year	Year	au	am	tt	ar	sa	amk	pa	ri	smu
term	Term	au	am	tt	ar	sa	amk	pa	ri	smu
semgpa	Semester gpa	au	am	tt	ar	sa	amk	pa	ri	smu
cgpa	Per semester cumulative gpa	au	am	tt	ar	sa	amk	pa	ri	smu
hrearn	Hours earned	au	am	tt	ar	sa	amk	pa	ri	smu
gpahrs	Hours used to calculate gpa		am							
term_major_dept	Department of college major	au	am	tt	ar	sa	amk	pa	ri	smu
term_major_field	Field of college major	au	am	tt	ar	sa	amk	pa	ri	smu

Note: Initials in the cell indicate that a variable is available for that institution.

VARIABLE DESCRIPTIONS

Variable Name	Variable Label
studentid	Student id

Description
Randomly generated identifier for each application.

Notes
The studentid variable may be used to merge College Transcript Data with associated Application Data.

Student identifiers are not unique across institutions. Therefore, if building a dataset that combines College Transcript Data from more than one institution, creation of an institutionid variable is recommended.

Variable Name	Variable Label
year	Year

Description
Calendar year of college transcript.

Variable Name	Variable Label
term	Term

Description
Academic term of college transcript.

Values	Value Labels
1	Spring
2	Summer I
3	Summer II
4	Fall
5	Spring intersession (Texas A&M Kingsville only)
6	Winter intersession (Texas A&M Kingsville only)

Variable Name	Variable Label
semgpa	Semester gpa

Values	Value Labels
1	0 – 0.5
2	0.6 – 0.6 (not available in Rice)
3	0.7 – 0.7 (not available in Rice)
4	0.8 – 0.8 (not available in Rice)
5	0.9 – 0.9 (not available in Rice)
6	1.0 – 1.0 (not available in Rice)
7	1.1 – 1.1

8	1.2 – 1.2	
9	1.3 – 1.3	
10	1.4 – 1.4	
11	1.5 – 1.5	
12	1.6 – 1.6	
13	1.7 – 1.7	
14	1.8 – 1.8	
15	1.9 – 1.9	
16	2.0 – 2.0	
17	2.1 – 2.1	
18	2.2 – 2.2	
19	2.3 – 2.3	
20	2.4 – 2.4	
21	2.5 – 2.5	
22	2.6 – 2.6	
23	2.7 – 2.7	
24	2.8 – 2.8	
25	2.9 – 2.9	
26	3.0 – 3.0	
27	3.1 – 3.1	
28	3.2 – 3.2	
29	3.3 – 3.3	
30	3.4 – 3.4	
31	3.5 – 3.5	
32	3.6 – 3.6	
33	3.7 – 3.7	
34	3.8 – 3.8	
35	3.9 – 3.9	
36	4.0 – 4.0	
37	4.1 – 4.1	(Rice, UT Arlington only)
38	4.2 – 4.2	(Rice, UT Arlington only)
39	4.3 – 4.3	(Rice, UT Arlington only)
40	4.4 – 4.4	(UT Arlington only)
41	4.5 – 4.5	(UT Arlington only)
42	4.6 – 4.6	(UT Arlington only)
43	4.7 – 4.7	(UT Arlington only)
44	4.8 – 4.8	(UT Arlington only)
45	4.9 – 4.9	(UT Arlington only)
46	5.0 – 5.0	(UT Arlington only)

Notes

Grade point averages are rounded to the nearest tenth, from 0 to 4.0 (0 to 4.3 at Rice, and 0 to 5.0 at UT Arlington). Because of small cell sizes, semester grade point averages for Rice between 0.5 and 1.0 are set to missing.

Variable Name	Variable Label
cgpa	Per semester cumulative gpa

Values	Value Labels
1	0 – 0.5 (not available in Rice)
2	0.6 – 0.6 (not available in Rice)
3	0.7 – 0.7 (not available in Rice)
4	0.8 – 0.8 (not available in Rice)
5	0.9 – 0.9 (not available in Rice)
6	1.0 – 1.0 (not available in Rice)
7	1.1 – 1.1
8	1.2 – 1.2
9	1.3 – 1.3
10	1.4 – 1.4
11	1.5 – 1.5
12	1.6 – 1.6
13	1.7 – 1.7
14	1.8 – 1.8
15	1.9 – 1.9
16	2.0 – 2.0
17	2.1 – 2.1
18	2.2 – 2.2
19	2.3 – 2.3
20	2.4 – 2.4
21	2.5 – 2.5
22	2.6 – 2.6
23	2.7 – 2.7
24	2.8 – 2.8
25	2.9 – 2.9
26	3.0 – 3.0
27	3.1 – 3.1
28	3.2 – 3.2
29	3.3 – 3.3
30	3.4 – 3.4
31	3.5 – 3.5
32	3.6 – 3.6
33	3.7 – 3.7
34	3.8 – 3.8
35	3.9 – 3.9
36	4.0 – 4.0
37	4.1 – 4.1 (Rice only)
38	4.2 – 4.2 (Rice only)
39	4.3 – 4.3 (Rice only)

Notes

Cumulative grade point averages are rounded to the nearest tenth, from 0 to 4.0 (0 to 4.3 at Rice). Because of small cell sizes, cumulative grade point averages for Rice in the categories below 1.1 are set to missing.

Variable Name	Variable Label
hrearn	Hours earned

Values	Value Labels
0	0 – 0
1	1 – 1
2	2 – 2
3	3 – 3
4	4 – 4
5	6 – 6
7	7 – 7
8	8 – 8
9	9 – 9
10	10 – 10
11	12 – 12
13	13 – 13
14	14 – 14
15	15 – 15
16	16 – 16
17	17 – 17
18	18 – 18
19	19 – 19
20	20 – 20
21	21 – 21
22	22 – 22
23	23 – 23
24	24 – 24
25	25 or more

Notes

The variable is available in whole hours from 0 to 24. More than 24 hours in a semester is collapsed into one group "25 or more" because of small cell sizes. Unusually high number of hours may be due to AP credits, transfer credits or completion of pending incompletes.

Variable Name	Variable Label
gpahrs	Hours used to calculate grade point average

Descriptions

Number of hours earned that contribute to calculation of semester and cumulative GPA.

Availability

Available only for Texas A&M.

Variable Name	Variable Label
term_major_dept	Department of college major

Notes

Enrollee major department may change between terms. The listing of all major departments is included in "Appendix B – Field of Major and Departments of Major by Field."

Variable Name	Variable Label
term_major_field	Field of college major

Notes

The values of the variable term_major_dept are grouped into fields in the variable term_major_field. For example, if an enrollee's term_major_dept is "Biology", then the term_major_field variable has the value "Natural/Physical Sciences".

The term_major_field variable provides an enrollee's actual field of major by term. This variable does not have the same meaning as the variable major_field in the Application Data. The major_field variable provides the field of each applicant's desired first choice major. A listing of departments and fields is included in "Appendix B – Field of Major and Departments of Major by Field."

MERGING COLLEGE TRANSCRIPT WITH APPLICATION DATA

Merging the College Transcript Data and Application Data enables researchers to track enrollees' academic progress. Seven of the universities have college transcripts for every student enrolled. The two exceptions are:

- *UT Arlington*: There are 4,473 applicants whose enrollee status is missing. The data does not include transcripts for these applicants.

- *Southern Methodist*: There are 1,380 enrollees for whom there are no college transcripts. These students were new admits who enrolled in 2005, but had not completed a term at the time the THEOP data collection period ended.

APPENDIX A – MERGING INSTRUCTIONS

Below are instructions for merging College Transcript Data and Application Data using the statistical program Stata. Although the instructions use UT Austin as an example, similar instructions apply to all nine institutions.

Sorting studentid

1. Save the college transcript file and the college application file in the same folder.
2. Sort the college transcript file and the college application file by `studentid`.

Merging College Transcript File with College Application File

3. Open the college transcript file, `theop_au_college_transcripts.dta`
4. Enter command:

   ```
   merge studentid using theop_au_college_applications.dta
   ```

5. Check the merge. The tabulation of `_merge` in UT Austin's case should look like the tabulation below:

   ```
    tab _merge

        _merge |      Freq.         Percent      Cum.
   ------------+-----------------------------------
             2 |    122,850        15.71        15.71
             3 |    659,102        84.29       100.00
   ------------+-----------------------------------
         Total |    781,952      100.00
   ```

   - _merge = = 2 represents values found only in the college application file. These are the 122,850 applicants who did not enroll at UT Austin and therefore have no college transcript information.
   - _merge = = 3 represents a match on studentid found in both the college transcript file and the college application file. These are the college transcripts of all students who chose to enroll. There are 87,156 students enrolled at UT Austin during the study period with 659,102 transcripts. Each student may have multiple transcripts – one for each semester enrolled.

6. Enter command:

   ```
   tab enroll _merge, m.
   ```

   This command checks that all enrollees are represented in the college transcript file and no non-enrollees have college transcripts.

APPENDIX B – FIELDS OF MAJOR AND DEPARTMENTS OF MAJOR BY FIELD

Fields in Bold, Majors within field indented below

Agriculture
Architecture
Business
- Accounting
- Business/Management
- Finance
- Marketing/Advertising
- Restaurant and Hotel
- Business, Other

Education
Engineering/Computer Science
- Engineering
- Computer/Information Science

Fine Arts
- Fashion/Interior Design
- Performing Arts
- Textiles
- Visual Arts
- Fine Arts, Other

General Studies
Health
- Allied Health
- Nursing
- Pre-Dentistry
- Pre-Medicine
- Pre-Pharmacy
- Pre-Veterinary Medicine
- Sports/Nutritional Sciences
- Health, Other

Humanities
- English/Literature
- Foreign Language
- Philosophy
- Regional Studies
- Religion
- Humanities, Other

Individualized/Interdisciplinary
Natural/Physical Sciences
- Biology
- Chemistry
- Environmental Science/Studies
- Geography
- Geology
- Mathematics
- Physics
- Zoology
- Natural/Physical Sciences, Other

Other
- Other
- Unknown

Social Sciences
- Anthropology
- Communications
- Criminal Justice
- Economics
- Government/Political Science/Pre-Law
- History
- Human/Child Development
- International Studies
- Journalism
- Policy Studies
- Psychology
- Sociology
- Urban Studies
- Social Sciences, Other

Social Work
Technical/Vocational
Undeclared