Report for Tennessee Higher Education Commission
Review and Analysis of SIS Data Files

Submitted by

George Malo

June 2011

Report for Tennessee Higher Education Commission
Review and Analysis of SIS Data Files

Executive Summary

A comprehensive analysis of five major Tennessee Higher Education Commission (THEC) student information system files was completed in May 2011. Five Student Information System (SIS) files were analyzed: 1) SIG file, 2) Graduate file, 3) Lottery file, 4) Term file, and 5) Term Credit file. Although records from years as early as 1994 were in the files, the analysis focused on records from fall 2004 through fall 2010, covering the period of time during which the Tennessee Education Lottery Scholarship program was in effect.

The purpose of the analysis was to determine the degree of accuracy of the data in the SIS files and to identify issues related to that data. Knowing the validity of the data and the related issues will help determine adjustments that may be needed in various reports. All records for each file were reviewed for identifying any issues with coding, whether within a particular field or across fields or files. The number of records for fields without codes or with improper coding was identified.

Although some varying degree of data noise was found, the analysis did not reveal any major concerns with those issues. In many cases the noise was minimal; and in areas where issues were found, adjustments can be made as recommended in the full report. The analysis also revealed a higher education data system that can be fully utilized because of its ability to provide data without violating the confidentiality of students. The structuring of the files independent of student social security numbers is a strong feature. Basing each record within the files on a unique, non-confidential identifier provides a robust data system for policymakers and researchers. The analysis did not find any problems with the identifier.

<u>Data Policy Recommendations</u>

1. Continue to meet periodically with boards.
2. Develop range and cross edits.
3. Use TSAC data for lottery reporting.
4. Revisit or review timelines.
5. Regularly review need for data elements or fields.

<u>Technical Recommendations</u>

1. Review/Clarify data element definitions
2. Back fill data fields when necessary.
3. Filter data when necessary.
4. Obtain data from other sources.

Report for Tennessee Higher Education Commission
Review and Analysis of SIS Data Files

A comprehensive analysis of five major Tennessee Higher Education Commission (THEC) student information system files was completed in May 2011. This report provides the results and recommendations from the analysis. The report is presented in three sections: 1) Methodology and Considerations, 2) Analysis of Files, and 3) Overall Summary and Recommendations.

Methodology and Considerations

Introduction

Five Student Information System (SIS) files were analyzed: 1) SIG file, 2) Graduate file, 3) Lottery file, 4) Term file, and 5) Term Credit file. Although records from years as early as 1994 were in the files, the analysis focused on records from fall 2004 through fall 2010. This time period for the analysis focus was chosen since fall 2004 was the beginning of the Tennessee lottery and several new data elements were also added to the graduation and term files during that period. Table 1 on the next page provides a general overview of each file.

While the years prior to 2004 may be useful for some long term trend analysis, the seven years from fall 2004 through fall 2010 provide sufficient data for determining trends and data patterns. These data also need to be as accurate as possible to provide valid trends and patterns. Thus, the purpose of the analysis reported in this paper was to determine the degree of accuracy of the data in the SIS files and to identify issues related to that data. Knowing the validity of the data and the related issues will help determine adjustments that may be needed in various reports.

The 2010-2011 THEC Public Institution Data Dictionary was used to identify the acceptable codes within each data field of the files. In addition, the dictionary defined the minimum edit checks for each variable (i.e., field). A copy of this dictionary is available from THEC.

Table 1: File Overview

File	Description	Number of Records	Number of Records 2004 - 2010	Number of Fields	Comments
SIG	Provides general demographics by student.	1,309,195	Not applicable since file is by student ID and not by date.	26	One record per student.
Graduate	Provides degree information by student.	400,872 (343,710 unique)	215,955 (194,050 unique)	19	One record per degree. Earliest award year in file is 1997.
Lottery	Provides information about students receiving the lottery scholarship.	539,137 (160,875 unique)	539,137 (160,875 unique)	14	One record per term by year. Earliest year in file is 2004, the year lottery began.
Term	Provides general information about student major, hours earned, etc.	6,388,541 (1,282,774 unique)	2,882,019 (667,163 unique)	16	One record per registration term by year. Earliest year in file is 1994.
Term Credit	Provides credit hour information about students.	6,388,510 (1,282,770 unique)	2,881,986 (667,158 unique)	58	One record per registration term by year. Earliest year in file is 1994.

Approach

Each data file was loaded into a database for analysis. Since the THEC ID was the only common element across all files, it was used as the index or key for joining the files. When appropriate to the analysis, multiple keys were also used to track records for individual students across multiple years and terms. Although the data were analyzed, for the most part, by individual file, joining the files also provided the ability to analyze the data across files. For example, the SIG file has a "lottery status" field (in-state or out-of-state values) and can be checked for field completion if the student is also in the Lottery file.

All records for each file were reviewed for identifying any issues with coding, whether within a particular field or across fields or files. The number of records for fields without codes or with improper coding was identified. Tables are provided for the various data fields within each file as an aggregate since providing tables by system sector and institution would result in an unmanageable number of table pages for the utility gleaned, especially when some tables displayed multiple fields.

Data Noise

Data noise is irrelevant or meaningless data that is usually a result of errors in the data set. As with most data systems in a social scientific setting, such as data collection around student information, some amount of noise is expected. Even after editing the data, some noise will remain. Because noise usually does not affect the main trends, one should decide how much noise is acceptable before conducting any analysis or producing reports from the data. If the noise is small or low level and does not have a significant impact on analysis or reporting, ignoring the noise can be an option. However, since the goal is to enhance any analysis or reporting as much as possible, data cleaning techniques that remove noise are generally the route taken. This data cleansing process can involve discarding the data causing the noise, adjusting the data to account for noise, or correcting the data, if there is practicality in doing so.

The tables presented in this report attempt to identify the degree of noise for the variable under review. This presentation will help in determining which avenue to take for adjusting any noise. Although coding errors can be identified, the analysis of the data files at this level is not able to always identify inaccurate data. For example, for the gender variable, one can determine if the coding was done correctly using one of the two defined codes, but one is unable to determine if that code used for the student accurately reflects his/her gender.

Analysis of Files

Introduction

The discussion in this section of the report provides an analysis of each field within a particular file. Whenever there was efficacy to look at data across fields or across files, the tables will reflect this interaction in the headings or in the brief comments about the tables. In some cases, an observation or recommendation will also be provided.

SIG File

The number of records reviewed was 1,309,195 records. The SIG file provides general demographics for each student enrolled in either public or private institutions. Since there is only one record per student, each record is unique.

THEC ID

All records had a student identification number. The student ID is unique with one record per ID. No problems with the ID were identified.

Observation: The ID cannot be tracked to any particular student without the master key held by THEC. Thus, student confidentiality has been maintained.

Gender

Gender coding does not present any data issues. Although one cannot verify accurate coding, there were no problems identified. This field had only 6 incorrect codes (values other than M and F) and 2 blanks.

Observation: The very small number of errors has no negative impact on data reporting and shows that the editing process for this field has been effective.

	Coded	**Incorrect**	**Blank**	**Totals**
Gender	1,309,187	6	2	1,309,195

Birth Year

The birth year provided for 3,173 records seemed suspect or were coded as “0”. This number accounts for 0.24% of the records and is within an acceptable “data noise” range. The table does not reflect a column for those born after 1929 since the SIG file contained students who may have been born after 1929. It was highly unlikely that students born prior to 1930 would be in the file.

Recommendation. Age calculations, such as average age of students, should be filtered to account for “0” coding and for suspect age ranges.

	Coded 0	**Born before 1920**	**Born 1920-1929**	**Totals**
Birth Year	242	53	2,878	3,173

Race

Race coding did not present any data issues. The very small number of errors has no negative impact on data reporting and shows that the editing process for this field has been effective. Race is dependent on students providing information and may change when a student switches institutions. The new race and ethnicity codes will need to be bumped against the student ID to update the SIG file to the new reporting requirements.

Observation. Although a student may report a different race when changing institutions, the SIG file is the appropriate place to collect race information because of the relational nature of the file to the other data bases.

Recommendation. As already planned by THEC, the new race field values should be according to the new reporting values rather than the data collection values. When new data are not available, the value should be the cross walk recommended by the National Center for Statistics (see http://nces.ed.gov/ipeds/reic/collecting_re.asp).

	Coded	Incorrect	Blank	Totals
Race	1,309,190	0	5	1,309,195

Citizenship and State Code Fields

Only 15 records were blank for the Citizenship field with seven of these records also not coded in the State Code field. Overall, 14,848 records were blank for the state code accounting for a 1.1% error. Of the blank records, 11,935 (80%) records had the permanent zip code completed. Thus, it should be possible to glean the state from the zip code, even though this is the permanent zip code. Also of those records that were blank, 95% of them were students born before 1985.

Observation. Error is within acceptable data noise range.

Recommendation. If state data are blank, one should back fill this field from the permanent zip field keeping in mind the permanent residence may be different than current residence.

Citizenship	State			Totals
	TN	Not TN	Blank	
US	1,132,208	118,563	14,045	1,264,816
Foreign Temp	1,357	29,505	645	31,508
Foreign Perm	11,469	1,237	150	12,856
Blank	7	1	7	15
Totals	1,145,041	149,306	14,848	1,309,195

Lottery Status

If a student is in the Lottery file, the lottery status should be coded in the SIG file. Of those that should have been coded in the SIG file, 1,033 (0.64%) were not coded.

Observation. Error is within acceptable data noise range.

Recommendation. For missing lottery status codes, one should back fill the SIG file from Lottery file.

	Student In Lottery File	Student Not In Lottery File	Totals
Status Coded	159,842	331,580	491,422
Status Not Coded	1,033	816,740	817,773
Totals	160,875	1,148,320	1,309,195

Zip Code

22% of the records for the ZIP code field were blank when Tennessee was coded as state. When considering residency code of 1, close to 1.3% of the ZIP code field was blank. *Observation.* When considering age, 991 (0.4%) students born in 1986 and later did not have a zip code when Tennessee was coded as a state. Thus, the problem tends to be with records of older students. Since emphasis is on more recent students for use of zip code, the data noise is very minimal.

	TN	Not TN	Total Blank
ZIP Blank	246,437	46,165	292,602
ZIP Blank or Bad when Residency = 1	11,337		

Residency

26% of values with state code of TN were not coded for residency. Of this 26%, approximately 99% were born prior to 1986 (1986 year was used since that was the year the first lottery students were born).
Recommendation. Residency and fee payment status should be reviewed with campuses to ensure everyone understands definitional differences and that data inconsistency exist.

Residency	**State**			**Totals**
	TN	**Not TN**	**Blank**	
1 (in-state)	843,294	7,829	324	851,447
2 (out of state)	4,326	88,360	178	92,864
3 (Foreign)	115	10,872	0	10,987
Blank	297,306	42,245	14,346	353,897
Totals	1,145,041	149,306	14,848	1,309,195

High School Grad Year

58% of the records did not provide the year the student graduated from high school. Some of those records missing high school graduation years are those of older students, so more recent high school students were checked (those born after 1986). 15% of the records were missing the high school graduation year for students who were born between 1986 and 1993. 41% of high school graduation year data was not provided for Tennesseans with in-state residency. 15% of high school graduation year data was not provided for Tennesseans with in-state residency who were born between 1986 and 1993.
Of the 23,962 missing high school graduation year records for Tennessee residents, 3,288 (13.7%) were lottery students
Recommendation. Populating high school data improved during the lottery years. Also some data are missing due to the time colleges receive information from high schools. Many institutions do not receive high school transcripts in time to populate this field when data files were due to THEC. Also, the older students tend not to have high school information on file. Institutions should be reminded to update their records when they do receive the high school information and should also update this record on subsequent file submissions to THEC. The field can then be checked for updates and back filled when necessary.

All Students

	High School Grad Year Provided	Blank	Total
All	549,201	759,994	1,309,195
Born ’86-‘93	231,794	42,454	274,248

Tennessee Residents

Condition	**Tennessee and Res = 1**		**Total**
	Year Coded	**Blank**	
All	497,806	345,488	843,294
Born ’86-‘93	140,226	23,962	164,188

Of the 23,962 missing high school graduation year records, 3,288 (13.7%) were lottery students

High School GPA

48% of high school GPA data (GED or HS) were not provided for Tennesseans with in-state residency. 21% of high school GPA data (GED or HS) were not provided for Tennesseans with in-state residency who were born between 1986 and 1993

Recommendation. As with high school graduation year, some data are missing due to the time colleges receive information from high schools and timeline for verifying GPA. Also, the older students tend not to have high school information on file. As with the previous field, recommend back filling when institutions receive high school data.

Condition	**Tennessee and Res = 1**		**Total**
	GPA Provided	**Blank**	
All	437,835	405,459	843,294
Born ’86-‘93	129,143	35,045	164,188

Of the 23,962 missing high school graduation year records, 6,521 (19%) were lottery students

High School Code

39% of the records for Tennessee residents did not provide the high school code. Some records missing the codes are those of older students, so more recent high school students were checked (those born after 1986). 15% of the records were missing the high school code for students who were born between 1986 and 1993.

Recommendation. Many institutions do not receive high school transcripts in time to populate this field when data files were due to THEC. Also, the older students tend not to have high school information on file. As with previous fields, recommend back filling whenever information is received at the institution.

Condition	**Tennessee and Res = 1**		**Total**
	High School Code Provided	**Blank**	
All	513,486	329,808	843,294
Born ’86-‘93	137,487	26,701	164,188

Of the 26,701 missing high school code records, 60 (0.2%) were lottery students

Curriculum Type

43.5% of the records for Tennessee residents did not provide the high school curriculum. More recent high school students were also checked (those born after 1986). 7% of the records were missing the curriculum type for students who were born between 1986 and 1993.

Recommendation. Many institutions do not receive high school transcripts in time to populate this field when data files were due to THEC. Also, the older students tend not to have high school information on file. Recommend back filling whenever information is received at the institution or populating as much high school information as possible from the XAP system at http://www.collegefortn.org.

Condition	**Tennessee and Res = 1**		**Total**
	Curriculum Type Provided	**Blank**	
All	476,566	366,728	843,294
Born '86-'93	137,700	26,488	164,188

ACT/SAT Composite

58% of ACT/SAT scores were not provided for Tennesseans with in-state residency. Some of those without scores are older students, so more recent high school students were checked (those born after 1986). 16% of ACT/SAT scores were not provided for Tennesseans with in-state residency who were born between 1986 and 1993.

Recommendation. Populating ACT/SAT data improved during the lottery years, however the percentage (16%) is still high for recent high school graduates. Recommend back filling by routinely referring to data files received by the Tennessee Student Assistance Corporation (TSAC).

Condition	**Tennessee and Res = 1**		**Total**
	ACT or SAT Provided	**Blank**	
All	356,623	486,671	843,294
Born '86-'93	137,487	26,701	164,188

Grad File

The following discussion about the graduation file is presented by each of the 19 fields in the file. The number of records reviewed was 215,955. Of these records, 194,050 had unique THEC IDs.

THEC ID

Although each student has a unique THECID, there are multiple records per student in the graduation file. As in all the files, student confidentiality has been maintained. No problems with the ID were identified.

System and Institution

No problems arose with the System and Institution fields with exception that ETSU had 334 records for MD degree under institution code 23 (ETSU) and 62 records for MD degree under institution code 88 (College of Medicine). The 62 records show all degrees awarded in 2010 while the 334 records show degrees awarded prior to 2010. The Report of Graduate instructions did not show a Quillen College of Medicine code until 2010. Prior to the 2010 year, all degrees were reported under the 23 institution code.

Although not the focus of this analysis report, it should be noted that there were 2408 records earlier than 2004 that had a system code of "8", Technical Institutes. This was an appropriate code prior to 2000 when Nashville became a community college.

All records had both a system and an institution identified.

Recommendation. Look at feasibility of moving institution code to 88 for all ETSU MD degrees prior to 2010. This will prevent having to remember to search on both institution codes.

Location

No issues or problems identified for this field. The off-campus awards matched approved locations for awarding degrees (e.g., APSU Fort Campbell, ETSU certificate and Associate degree in health sciences, UT, Volunteer State Livingston).

Award Term

No issues or problems identified for this field. All records had an appropriate code.

Award Year

No issues or problems identified for this field. All records had an appropriate year.

Completion of Requirements Term

No issues or problems identified for this field. All records had an appropriate year.

Completion of Requirements Year

604 records had the completion year as blank and 15 records had incorrect year data (e.g., 2088).

Recommendation. Estimate completion year with award year. It should be noted that in the majority of cases, the year in which the degree is awarded and the year for completion of requirements are the same. This element will differ from award year only if no formal graduation ceremonies were held at the end of the term in which requirements were completed. For example, in some cases a student may complete his/her work in the late summer or fall, but the actual award is given in the spring if that is the only time graduation ceremonies are held. These time periods are still close enough to estimate for data purposes.

Degree

The number of records with no degree coding was 63,912 or 29.6%.

Recommendation. Since the degree is the alphanumeric value (e.g., BA) and the award level is the code (e.g., 25 for a BA), obtain the degree by using award level codes.

Award Level

No issues or problems identified for this field. All records had an award level.

Award Degree

The number of records with no award degree coding was 12,300 or 5.7%. All records with missing award degree codes also did not have coding for the degree field.

Recommendation. Since the award degree is the join of degree and award level (e.g. 2.5BA), estimate degree by using the award level.

First Major and other Major fields

The number of records with no first major coding was 30. Of those, 29 had the additional major field coded (these did not have second major coded either).

Recommendation. Estimate major by looking at additional major field. For the 1 record having no coding, do not include it in degree analysis or reporting since one record will have no impact if eliminated.

Total Credit Hours Attempted

The number of records with no total credit hours attempted was 214,817, or almost all the records. Of those, 52,974 had the total credit hours earned coded. Those with coding were for only the award years 2007 and 2008.

Recommendation. This field is fairly new, but may be problematic for reporting. Review the need for this field and its importance in graduation reports.

Total Credit Earned and GPA

The number of records with no total credit earned coding was 161,843 or 75%. Although the percentages were high, the collection of these data is new. The GPA provided did not show any problems, however the credit hours provided were high in many cases. The credit hours were not consistent since it seemed some were for the degree only and others were cumulative for all college work. For example, a student attending a community and graduating with an associate's degree after receiving a bachelor's degree may show credits for all degrees.

Recommendation. Estimate hours for the first collections using cumulative term credit, when possible. Meet with the institutions to discuss reporting and edit checks of just the hours cumulated for the specific degree and not all hours cumulated when more than one degree has been received or additional coursework has been completed after receiving a degree.

	Provided	**Blank**	**Percent Blank**	**Totals**
Credit Hours	54,112	161,843	75%	215,955
Final GPA	75,378	140,577	65%	215,955

Lottery Amount Received and Lottery GPA

All records were blank. These are new fields.

Recommendation. Get data from TSAC. Consider also the need for these fields on the graduation file and whether the lottery file can provide the same information.

Lottery File

The following discussion about the Lottery file is presented by each field in the file. The number of records reviewed was 539,137. Of these records, 160,875 had unique THEC IDs. It should be noted that raw numbers rather than percentages are reported by institutional sector. Thus, the incidence rate of error should not be determined by the size of the numbers. Likewise, some of the incidences of missing numbers and other errors are redundant due to duplicate records for students. Thus, the incidence rate of error would be lower if considered on a per student basis.

THEC ID

Although each student has a unique THECID, there are multiple student records in the Lottery file. As in all the files, student confidentiality has been maintained. No problems with the ID were identified.

System and Institution

No issues or problems arose with the System and Institution fields. All records had both a system and an institution identified.

Registration Term and Registration Year

No issues or problems arose with registration term or year fields. All records had both a registration term and a registration year.

Initial Lottery Year

Of the 539,137 records, 21,012 (3.9%) have missing Initial Lottery Year information. Some of the incidences of missing numbers are redundant due to duplicate records for students. *Recommendation.* Discuss how these data are used and determine if necessary to back fill or if counting the number of registration years will be sufficient for missing data needs.

	TBR Univ	**Community Colleges**	**UT**	**Private**	**Totals**
Missing	767	8,453	4,496	7,296	21,012

Cumulative Credit Hours

Of the 539,137 records, 201,197 (37%) have missing Cumulative Credit Hours information. Some of the incidences of missing information may be due to the duplicate records for some students and the first term of the lottery, but this percentage is still quite high. *Recommendation.* Obtain data from TSAC or update with most recent cumulative hours.

	TBR Univ	**Community Colleges**	**UT**	**Private**	**Totals**
Missing	88,812	53,360	39,538	19,487	201,197

Lottery GPA

Of the 539,137 records, 177,085 (33%) have missing lottery GPA information. Some of the incidence of missing information is due to the duplicate records for some students and first term data. *Recommendation.* Obtain data from TSAC.

	TBR Univ	Community Colleges	UT	Private	Totals
Missing	61,678	44,177	39,839	31,391	177,085

Lottery Scholarship Type

Of the 539,137 records, 19 records have codes not in the data element dictionary. See table on next page for more detail.

Recommendation. No issues have surfaced with this field.

Term Lottery Amount

Of the 539,137 records, some lottery amounts are absent for the Z code while others occur when scholarships were shown to be lost. See table on next page for more detail. Dollar amounts in some cases were not entered correctly, especially with understood 00. For example, $1500 was entered as $1500 rather than $150000 with the last two zeros understood as decimal. Thus, $1500 would come out as $15. Also see attached table showing range of dollars.

Recommendation. The fact that some lottery dollar amounts do not agree with lost scholarship reasons and that some lottery dollar amounts are small or non-existent will need to be reviewed. Obtain the term lottery amount from TSAC.

Lost Scholarship Reason

Of the 539,137 records, some lottery amounts are absent for the Z code while others occur when scholarships were shown to be lost. See table on next page for more detail.

Recommendation. If a student shows a lost scholarship reason, one would expect the term lottery amount field to be blank. However, many records had data in the dollar amount field when a scholarship was lost. Similarly, when the student was shown to receive the lottery scholarship, the dollar amount field was blank or a very small amount (it should be noted that this could happen if a student turned down the scholarship or other aid sufficiently covered costs). The fact that some lottery dollar amounts are not consistent with the scholarship type or lost scholarship reason needs to be reviewed.

The Lottery file had several issues that need to be reviewed. The recommendation is that TSAC data be used for reporting or, at the very least, for validation. Reporting of lottery scholarship in the past was also at a time when schools did not always know who was receiving the lottery. Thus, many of the issues indicated above were due to the statutory necessitated collection timeline, which is no fault of THEC or the Boards since the lottery report was due to the legislature in mid-January.

Lottery File
Lost Scholarship By System

					Schl Type and Amt = 0								Schl Type and Amt > 0										
System	Lostscholr	Amt=0	Amt>0	Total	1	2	3	4	5	6	7	Total	0	1	2	3	4	5	6	7	8	9	Total
1	Blank	3,006	16	3,022	2,052	70	800	84	0	0	0	3,006	0	2	0	13	0	1	0	0	0	0	16
1	A	49,769	529	50,298	35,253	1,554	12,376	586	0	0	0	49,769	0	300	4	220	5	0	0	0	0	0	529
1	B	3,130	18	3,148	2,267	223	640	0	0	0	0	3,130	0	14	0	4	0	0	0	0	0	0	18
1	C	235	2	237	0	0	0	235	0	0	0	235	0	0	0	1	1	0	0	0	0	0	2
1	D	2,809	156	2,965	1787	81	788	153	0	0	0	2,809	0	105	3	48	0	0	0	0	0	0	156
1	E	973	45	1,018	696	37	211	29	0	0	0	973	0	25	3	17	0	0	0	0	0	0	45
1	F	851	7	858	595	82	174	0	0	0	0	851	0	3	0	4	0	0	0	0	0	0	7
1	G	12	0	12	0	0	12	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0
1	O	3,479	401	3,880	2,759	112	554	27	0	27	0	3,479	0	218	18	162	3	0	0	0	0	0	401
1	U	9	0	9	6	0	3	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0
1	X	544	0	544	346	10	99	19	0	70	0	544	0	0	0	0	0	0	0	0	0	0	0
1	Z	4,310	160,429	164,739	2,593	215	1,157	314	0	31	0	4,310	0	111,354	9,114	39,127	813		15	5	0	1	160,429
2	0	2	4	6	2	0	0	0	0	0	0	2	3	1	0	0	0	0	0	0	0	0	4
2	1	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2
2	Blank	5,062	97	5,159	2,985	10	1,266	229	0	571	1	5,062	0	90	2	3	1	1	0	0	0	0	97
2	A	22,385	606	22,991	16,321	188	5,472	395	0	6	3	22,385	0	412		191	3	0	0	0	0	0	606
2	B	50	1	51	41	3	6	0	0	0	0	50	0	1	0	0	0	0	0	0	0	0	1
2	C	69	1	70	0	0	0	69	0	0	0	69	0	0	0	1	0	0	0	0	0	0	1
2	D	3,524	109	3,633	2,317	22	1,043	140	0	2		3,524	0	75	0	34	0	0	0	0	0	0	109
2	E	1,193	142	1,335	955	7	202	27	0	2	0	1,193	0	88	3	51	0	0	0	0	0	0	142
2	F	4	0	4	3	0	1	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
2	G	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
2	O	5,330	35	5,365	4,460	58	509	22	0	281	0	5,330	0	19	0	16	0	0	0	0	0	0	35
2	U	8	0	8	6	0	2	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0
2	X	4,575	0	4,575	1,446	3	438	67	0	2,621	0	4,575	0	0	0	0	0	0	0	0	0	0	0
2	Z	10,259	68,011	78,270	4,437	58	1,892	285	0	3,587	0	10,259	0	46,652	304	20,221	763		59	10	1	1	68,011
3	Blank	1	42	43	0	0	0	0	0	1	0	1	0	36	0	6	0	0	0	0	0	0	42
3	A	10	49	59	7	0	3	0	0	0	0	10	0	34	1	14	0	0	0	0	0	0	49
3	B	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
3	C	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2
3	D	9	13	22	7	0	2	0	0	0	0	9	0	11	0	2	0	0	0	0	0	0	13
3	E	18	0	18	11	2	5	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0
3	F	19	1	20	19	0	0	0	0	0	0	19	0	1	0	0	0	0	0	0	0	0	1
3	O	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
3	Z	240	128,255	128,495	224	4	12	0	0	0	0	240	0	87,296	14,764	21,146	346		4,464	1	224	14	128,255
5	Blank	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	A	81	324	405	63	3	15	0	0	0	0	81	0	258	2	63	0	0	0	0	1	0	324
5	B	1	1	2	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1
5	D	0	3	3	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	3
5	E	11	13	24	5	4	2	0	0	0	0	11	0	6	2	5	0	0	0	0	0	0	13
5	O	3	14	17	2	1	0	0	0	0	0	3	0	10	2	2	0	0	0	0	0	0	14
5	Z	8,338	49,486	57,824	5,127	912	1,856	44	0	399	0	8,338	0	30,301	7,283	10,368	144		1,146	4	240	0	49,486
TOTALS		130,323	408,814	539,137	86,794	3,659	29,543	2,725	0	7,598	4	130,323	3	277,313	31,505	91,722	2,081	4	5,684	20	466	16	408,814

Lottery File
Lost Scholarship by Value

					Schl Type and Amt = 0								Schl Type and Amt > 0										
System	Lostscholr	Amt=0	Amt>0	Total	1	2	3	4	5	6	7	Total	0	1	2	3	4	5	6	7	8	9	Total
1	Blank	3,006	16	3,022	2,052	70	800	84	0	0	0	3,006	0	2	0	13	0	1	0	0	0	0	16
2	Blank	5,062	97	5,159	2,985	10	1,266	229	0	571	1	5,062	0	90	2	3	1	1	0	0	0	0	97
3	Blank	1	42	43	0	0	0	0	0	1	0	1	0	36	0	6	0	0	0	0	0	0	42
5	Blank	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
2	0	2	4	6	2	0	0	0	0	0	0	2	3	1	0	0	0	0	0	0	0	0	4
2	1	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2
1	A	49,769	529	50,298	35,253	1,554	12,376	586	0	0	0	49,769	0	300	4	220	5	0	0	0	0	0	529
2	A	22,385	606	22,991	16,321	188	5,472	395	0	6	3	22,385	0	412		191	3	0	0	0	0	0	606
3	A	10	49	59	7	0	3	0	0	0	0	10	0	34	1	14	0	0	0	0	0	0	49
5	A	81	324	405	63	3	15	0	0	0	0	81	0	258	2	63	0	0	0	0	1	0	324
1	B	3,130	18	3,148	2,267	223	640	0	0	0	0	3,130	0	14	0	4	0	0	0	0	0	0	18
2	B	50	1	51	41	3	6	0	0	0	0	50	0	1	0	0	0	0	0	0	0	0	1
3	B	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	B	1	1	2	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1
1	C	235	2	237	0	0	0	235	0	0	0	235	0	0	0	1	1	0	0	0	0	0	2
2	C	69	1	70	0	0	0	69	0	0	0	69	0	0	0	1	0	0	0	0	0	0	1
3	C	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2
1	D	2,809	156	2,965	1787	81	788	153	0	0	0	2,809	0	105	3	48	0	0	0	0	0	0	156
2	D	3,524	109	3,633	2,317	22	1,043	140	0	2		3,524	0	75	0	34	0	0	0	0	0	0	109
3	D	9	13	22	7	0	2	0	0	0	0	9	0	11	0	2	0	0	0	0	0	0	13
5	D	0	3	3	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	3
1	E	973	45	1,018	696	37	211	29	0	0	0	973	0	25	3	17	0	0	0	0	0	0	45
2	E	1,193	142	1,335	955	7	202	27	0	2	0	1,193	0	88	3	51	0	0	0	0	0	0	142
3	E	18	0	18	11	2	5	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0
5	E	11	13	24	5	4	2	0	0	0	0	11	0	6	2	5	0	0	0	0	0	0	13
1	F	851	7	858	595	82	174	0	0	0	0	851	0	3	0	4	0	0	0	0	0	0	7
2	F	4	0	4	3	0	1	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
3	F	19	1	20	19	0	0	0	0	0	0	19	0	1	0	0	0	0	0	0	0	0	1
1	G	12	0	12	0	0	12	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0
2	G	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
1	O	3,479	401	3,880	2,759	112	554	27	0	27	0	3,479	0	218	18	162	3	0	0	0	0	0	401
2	O	5,330	35	5,365	4,460	58	509	22	0	281	0	5,330	0	19	0	16	0	0	0	0	0	0	35
3	O	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	O	3	14	17	2	1	0	0	0	0	0	3	0	10	2	2	0	0	0	0	0	0	14
1	U	9	0	9	6	0	3	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0
2	U	8	0	8	6	0	2	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0
1	X	544	0	544	346	10	99	19	0	70	0	544	0	0	0	0	0	0	0	0	0	0	0
2	X	4,575	0	4,575	1,446	3	438	67	0	2,621	0	4,575	0	0	0	0	0	0	0	0	0	0	0
1	Z	4,310	160,429	164,739	2,593	215	1,157	314	0	31	0	4,310	0	111,354	9,114	39,127	813		15	5	0	1	160,429
2	Z	10,259	68,011	78,270	4,437	58	1,892	285	0	3,587	0	10,259	0	46,652	304	20,221	763		59	10	1	1	68,011
3	Z	240	128,255	128,495	224	4	12	0	0	0	0	240	0	87,296	14,764	21,146	346		4,464	1	224	14	128,255
5	Z	8,338	49,486	57,824	5,127	912	1,856	44	0	399	0	8,338	0	30,301	7,283	10,368	144		1,146	4	240	0	49,486
TOTALS		130,323	408,814	539,137	86,794	3,659	29,543	2,725	0	7,598	4	130,323	3	277,313	31,505	91,722	2,081	4	5,684	20	466	16	408,814

Distribution of Lottery Amounts

System	Termlotamt $ Range	Count	System	Termlotamt $ Range	Count
1	0	69,127	3	0	299
1	1	14,048	3	1	0
1	100	11	3	100	10
1	101-1499	1,501	3	101-1499	5,698
1	1,500	13,413	3	1,500	9,678
1	1501-1999	32,752	3	1501-1999	30,982
1	2,000	61,883	3	2,000	47,745
1	2001-2499	6,073	3	2001-2499	8,706
1	2,500	4,471	3	2,500	8,483
1	2501-2999	27,439	3	2501-2999	15,625
1	3,000	2	3	3,000	1,008
1	3001-3499	1	3	3001-3499	17
1	3,500	4	3	3,500	4
1	3501-3999	1	3	3501-3999	12
1	4,000	0	3	4,000	386
1	4001-4999	1	3	4001-4999	2
1	5,000	2	3	5,000	6
1	0.50	1	3	0.50	0
2	0	52,462	5	0	8,435
2	1	6,502	5	1	0
2	100	18	5	100	4,285
2	101-1499	48,165	5	101-1499	1,698
2	1,500	172	5	1,500	3,231
2	1501-1999	14,136	5	1501-1999	9,168
2	2,000	5	5	2,000	15,425
2	2001-2499	3	5	2001-2499	2,925
2	2,500	1	5	2,500	2,341
2	2501-2999	5	5	2501-2999	6,823
2	3,000	0	5	3,000	0
2	3001-3499	0	5	3001-3499	278
2	3,500	0	5	3,500	1
2	3501-3999	0	5	3501-3999	191
2	4,000	0	5	4,000	555
2	4001-4999	0	5	4001-4999	1,111
2	5,000	0	5	5,000	1,809
2	0.50	1	5	0.50	0

Term File

The following discussion about the Term file is presented by each field in the file. The number of records reviewed was 2,882,019. The number of records with unique student IDs was 667,163. It should be noted that raw numbers rather than percentages are reported by institutional sector. Thus, the incidence rate of error should not be determined by the size of the numbers. Likewise, some of the incidences of missing numbers and other errors are redundant due to duplicate records for students. Thus, the incidence rate of error would be lower if considered on a per student basis.

THEC ID

Although each student has a unique THECID, there are multiple student records in the Term file. The ID cannot be tracked to any particular student without the master key held by THEC. Thus, student confidentiality has been maintained. No problems with the ID were identified.

System and Institution

No issues or problems arose with the System and Institution fields. All records had both a system and an institution identified.

Registration Type

There are 6 codes for registration type. An error will occur if the registration type is not one of these codes (i.e., 1-6). Of the 2,882,019 records, 4,733 have missing registration type codes. The codes not blank were one of the 6 defined codes.

	TBR Univ	**Community Colleges**	**UT**	**Private**	**Totals**
Blank for Reg Type	7 (1 university)	41 (2 comm colleges)	2 (1 university)	4,683 (several inst)	4,733

Previous Registration Type

There are 4 codes for registration type. An error will occur if the registration type is not one of these codes (i.e., 1-4). Of the 2,882,019 records, 24,096 records have missing previous registration type codes. Except for 2 records, the codes not blank were one of the 4 defined codes.

	TBR Univ	**Community Colleges**	**UT**	**Private**	**Totals**
Blank for Prev Reg Type	7 (1 university in 1 yr only)	44 (2 comm colleges in 1 year only)	1 (1 university)	24,044 (several institutions over several yrs)	24,096

Registration Year and Registration Term

No issues or problems arose with registration year or term fields. All records had both a registration year and a registration term with all records within the defined codes.

Student Level

Except for 850 records that were blank, the student level field contained codes within the range of values defined for that field. Because only 4 institutions were affected and mostly within specific years, coding according to defined values was not considered a problem. (Student level crosschecks will be presented with major field section).

	System Code	Institution Code	Number Blanks	Comment
Student Level	3	86	1	----
	5	26	2	Transfer students
	5	41	68	All in 2004 only
	5	74	779	2009 only

Transfer Institution

A registration code of 3 indicates a transfer student. The number of records with a registration code of 3 was 156,622. Of the records indicating a transfer student (or undergrad institution for medical schools), 3,571 (2.3%) had missing values for identifying the transfer institution. The majority of the missing values were at the medical schools (86% of TBR missing codes were at the med school and 55% of UT were at the med school).

	TBR Univ	Community Colleges	UT	Private	Totals
Missing Transfer Inst Code	1,458 (1,253 at med school)	1,080	944 (522 at med school)	88	3,571 (2.3%)

Student Major

The student major field contains the academic major of the student as identified through the 10-digit code from the NCES Classification of Instructional Programs. Undeclared majors for undergraduates are identified through a code of "U" and special students are identified through a code of "N." The table below summarizes the number of records with various student major codes. Although it is difficult to check the accuracy of the major as to whether the proper major code was used, codes that are blank or less than ten digits are considered errors.

Major Code	TBR Univ	Comm Colleges	UT	Private	Totals
N	59,562	186,445	25,927	525	272,459
U	118,944	28,268	136,594	6,305	290,111
10 Digit	957,917	862,948	438,162	45,110	2,304,137
Blank	0	2	8,982	2,501	11,485
< 10 digit or zeros			1	3,826	3,827
TOTALS	1,136,423	1,077,663	609,666	58,267	2,882,019

The major field also interacts with the student level. For example, if the major field is equal to the undeclared major code, "U", then the student level must be one of the undergraduate level codes 1 through 4. The table shows some of the major field and student level interactions. The rows with asterisks indicate coding errors.

Major Code	**Student Level**	**TBR Univ**	**Comm Colleges**	**UT**	**Private**	**Totals**
U	**1, 2, 3, or 4**	118,913	27,904	136,586	5,947	289,350
U**	**Not 1-4****	31	364	8	358	761
N	**6, 10, or 40**	59,517	186,416	25,923	525	272,381
N**	**Not 6, 10, or 40****	45	29	4	0	78
TOTALS		178,506	214,713	162,521	6,830	562,570

Cumulative Credit Earned and Home GPA

There were no blank records for cumulative credit hours earned. The number of blank records for home GPA was 512,944 or 18%.

	Provided	**Blank**	**Percent Blank**	**Totals**
Credit Hours	2,882,019	0	0%	2,882,019
Home GPA	2,369,075	512,944	18%	2,882,019

Lottery Amount and Lottery Hours

The number of records showing lottery dollar amounts was 121,008. The number of records showing lottery hours was 261,322. Considering both fields together, 75,004 records show a lottery dollar amount when the lottery hours field also showed a value greater than zero. When considering unique records with a lottery amount shown, the Term file showed 65,453 records while the Lottery File showed 149,900 records.

	Values Shown by Field	**Values in Both Fields**	**Values in One Field Only**
Lottery Amount	121,008	75,004	46,004
Lottery Hours	261,322		186,318

Lottery GPA

The number of Lottery GPA records showing no value when both the Lottery dollar amount and Lottery hours had values was 8,267 records. The number of Lottery GPA records showing no value when the Lottery dollar amount field had values was 53,061 records.

Term Credit File

The following discussion about the Term file is presented by each field in the file. The number of records reviewed was 2,881,986. The number of records with unique student IDs was 667,158. It should be noted that raw numbers rather than percentages are reported by institutional sector. Thus, the incidence rate of error should not be determined by the size of the numbers. Likewise, some of the incidences of missing numbers and other errors are redundant due to duplicate records for students. Thus, the incidence rate of error would be lower if considered on a per student basis.

THEC ID

Although each student has a unique THECID, there are multiple student records in the Term file. The ID cannot be tracked to any particular student without the master key held by THEC. Thus, student confidentiality has been maintained. No problems with the ID were identified.

System and Institution

No issues or problems arose with the System and Institution fields. All records had both a system and an institution identified.

Registration Year and Registration Term

No issues or problems arose with registration year or term fields. All records had both a registration year and a registration term with all records within the defined codes.

Credit Type

No issues or problems arose with the credit type. All records had a credit type for Credit Type 1 that was within the defined values.

Delivery Type

Since this field was fairly new, records were checked for term years 2006 and beyond. Of the 6,388,510 records in the Term Credit file, those with term years of 2006 and beyond equal 2,200,317 records. Table below shows the number of missing records and coded records for the delivery type for each of these years. Those that were coded were within the codes expected. Only delivery type 1 is shown in the tables below. For the other delivery fields, the coding was as expected.

	TBR Univ	**Community Colleges**	**UT**	**Private****	**Totals**
Missing 2006	493	0	0	9,272	9,765
Missing 2007	261	15	0	9,134	9,410
Missing 2008	3	1	0	11,296	11,300
Missing 2009	0	1	0	12,416	12,417

	TBR Univ	Community Colleges	UT	Private	Totals
Coded 2006	190,106	174,416	100,790	0	465,312
Coded 2007	193,517	176,910	102,190	0	472,617
Coded 2008	194,393	181,548	107,210	0	483,151
Coded 2009	200,512	203,240	111,537	0	515,289
Coded 2010*	85,301	89,455	46,300	0	221,056

* Term 1 only. ** Private did not have to report delivery type during the term years.

Term Credit Hours and Total Credit Hours

Problems were found in two areas: Term credit hours greater than 24 (with many greater than 30) and records entered incorrectly. These records entered incorrectly were entered as 2 digits rather than the 4 digits with the right two digits understood as decimal. As a result, a student with 15 credit hours would be calculated as 0.15 credits.

Problem	TBR Univ	Comm Colleges	UT	Private
Credit Hours > 24	1,031	78	4,867	8
Comments	786 records were med school		4,727 were med school	
Entered Incorrectly	0	0	0	14,830
Comments				Resulted in 8,692 records showing total credit hours less than 1

Fee Paying Status

There are 20 codes for fee paying status with most of the records showing either a 1(in-state) or 2 (out of state). An error will occur if the fee paying status is not one of these codes. Of the 6,388,510 records, 61,505 have missing fee paying status codes. It should be noted that independent colleges do not report fee paying status and account for 58,234 records having no value in this field. Thus, 3,271 public institutions had missing fee paying status codes. The codes not blank were one of the defined codes.

	1 or 2 Code	Code other than 1 or 2	Records not Coded	Totals
Fee Paying	2,503,546	316,935	61,505 (includes privates)	2,881,986

	TBR Univ	Community Colleges	UT	Private	Blank Totals
Blank for Fee Paying	914	62	2,295 (2,282 records were in 2005 term for inst 30)	58,234 (this field not collected for private schools)	61,505

Overall Summary and Recommendations

Conclusions

This section of the report provides a summary of the analysis and offers recommendations for consideration. In general, the data for the years 2004-2010 were much cleaner than the earlier years provided in the files. This fact perhaps can be attributed to improved data editing and institutional collection techniques over the past several years. Although some varying degree of data noise was found, the analysis did not reveal any major concerns with that noise. In many cases the noise was minimal; and in areas where issues were found, adjustments can be made as recommended in this section of the report. Some of the identified issues impacting data quality may also be resolved as THEC goes to the new end of term data file collection.

The analysis revealed a higher education data system that can be fully utilized because of its ability to provide data without violating the confidentiality of students. The structure of the files with data analysis and tracking not dependent on student social security numbers is a strong feature. Basing each record within the files on a unique, non-confidential identification provides a robust data system for policymakers and researchers. The analysis did not find any problems with the identifier.

The analysis also revealed some areas that can be addressed internally and with the board and institutional staffs who provide the data. In some cases, the issues may be resolved through data collection timeline adjustments. In other cases, the issues can be resolved by addressing such things as more clarity in data element definitions, enhanced editing procedures, or identifying the best data source.

Suggested Areas for Review

The analysis found missing data within the lottery files. Perhaps due to the data collection timeline required of THEC for the legislative report and the fact that students can still apply for the lottery after classes begin, complete lottery data were not available for some students. The changing lottery rules in the beginning years also may have impacted data collection, especially in the consistency of data collection. A more stable source for the lottery data may be TSAC.

The graduate file should also be reviewed, especially with respect to the credit hours counted for graduation and the grade point averages. These elements are fairly new to the graduate file and may rectify some of the concerns as the elements mature. However, the hours for graduation were high in many cases (and in some, were below requirements) and should be reviewed. The graduation GPA showed some inconsistencies and in many cases was missing, perhaps due to the newness of the data elements.

In the lottery, term, and term credit files, the credit hours were not always consistent across files. For example, cumulative credit hours were not always sequential or were less than previous term. Discussions with the boards for ways of improving quality with credit hours may be beneficial as well as instituting edits that look at credit hours across files.

Data Policy Recommendations

1. Continue to meet periodically with boards.

   Meeting on a regular basis with board staff (UT, TBR, Independents) to discuss and resolve data issues is important to ensuring the best quality data. At times, K-12 data staff and appropriate others (e.g. business partners) should also meet with higher education staff to focus on the data pipeline, data use, and issues. Some areas that may be agenda items include agreements on definitions or clarification of data elements (e.g., residency coding, grade point average methodologies), problems with data collection (e.g., credit hours, graduate hours), data noise tolerances, timelines, new elements, or institutional data edits.

2. Develop range and cross edits.

   Many of the current edits just check for values. Develop range or cross file edits with the boards that campuses can also use. For example, cumulative hours from the term files may be an estimated cross check or range edit against hours at graduation.

3. Use TSAC data for lottery reporting.

   Consider the feasibility of using the TSAC data base for lottery reporting and validation (part of the editing process) of institutional files. These data may be timelier and should provide a more accurate picture at the time that lottery information is first needed.

4. Revisit or review timelines.

   Continue to review timelines with the boards, especially in light of legislation and time needed for enhanced edit checks. Review missing data to see if current timelines have played a role in the ability to collect that data when a file is due. In some cases it may be necessary to suggest a change in legislative reporting dates.

5. Regularly review need for data elements or fields.

   Continue the practice of regularly reviewing the need for new data elements and the utility of keeping current data elements. In light of new or changed policies, legislation, and Federal requirements, this practice becomes even more important to data collection and reporting.

Technical Recommendations

1. Review/Clarify data element definitions

   Periodically conduct checks (perhaps during meetings with the boards and the boards with their institutions) for clarification of data element definitions and consistency with policies. With the number of new IR personnel over time, it becomes even more important that all data providers have a clear understanding of data definitions.

2. Back fill data fields when necessary.

   In cases where adjustments for data noise are not feasible, it may be necessary to back fill missing or incorrect data. For example, it may be possible that some of the missing data in the lottery file can be back filled through the TSAC lottery database. High school data can be back filled when data are available.

3. Filter data when necessary.

   In cases where data are missing, it may be necessary to filter the data for reporting. For example, when reporting average age of students, age calculations should be filtered to account for "0" coding and for suspect age ranges. In some cases, ignoring suspect records in calculations may be appropriate.

4. Obtain data from other sources.

   In cases where data are missing, one may be able to estimate or interpolate these data from other sources. For example, high school information provided from the XAP system at http://www.collegefortn.org may be available for populating fields that have missing information.