Texas Higher Education Opportunity Project (THEOP) Baseline Survey

Methodology Report

August 22, 2003

TABLE OF CONTENTS

Introduction **4**

Survey Objectives **4**

Sample Design **5**

Frame. 5

Stratification. 6

Two stage probability sample. 7

Survey Instruments **7**

Field Methods **8**

District cooperation. 8

High school cooperation. 9

Logistical arrangements. 9

Survey Administration. 10

Exceptions. 11

Field periods. 12

Results **12**

School Participation. 12

Mail survey results. 13

In-class survey results. 13

Overall Results. 13

Data Processing **14**

Receipt Control. 14

Bursting. 14

Storing. 14

Scanning. 15

Verification. 15

Coding. 15

Data cleaning. 15

Appending Administrative Data. 15

Archiving survey images. 16

Analysis weights. 16

Sampling errors. 17

Appendix A **19**

LIST OF TABLES

Table 1: Eleven Principal Strata Used in the Texas Higher Education Opportunity Project Baseline Sample of Public High Schools 6

Table 2: Disposition of Schools in the Texas Higher Education Opportunity Project 12

Table 3: Texas Higher Education Opportunity Project Baseline Survey Student Level Response Rates for the Mail Component 13

Table 4: Texas Higher Education Opportunity Project Baseline Survey In-Class Student Level Response Rates 13

Table 5: Texas Higher Education Opportunity Project Baseline Survey Overall Student Level Response Rates 14

Table A1: Average Design Effects for Seniors by Race-Ethnic Group and Instrument Section 19

Table A2: Average Design Effects for Seniors by Race-Ethnic Group and Instrument Section 19

Table A3: Square Roots of Average DEFFs for SENIORS by Race-Ethnic Group and Instrument Section 19

Table A4: Square Roots of Average DEFFs for SOPHOMORES by Race-Ethnic Group and Instrument Section 20

A Methodological Report on the Texas Higher Education Opportunity Project (THEOP) Baseline Survey

Introduction

The purpose of this report is to document the design, methodology and outcomes of the Texas Higher Education Opportunity Project Baseline Survey. This report is divided into five sections. The first section describes the population of inference and the survey objectives. The second section describes the sample design. This is followed by sections describing the instrument development process and the field protocols. The results of the field effort are described in the fourth section. The final section is devoted to data processing and the creation of analytic weights.

The Baseline survey was conducted in spring of 2002 in the State of Texas. The survey population consisted of all seniors and sophomores attending public high schools in the state of Texas. Exclusions included students in charter schools and students in special education classes (and special education schools) and schools with fewer than 10 seniors (using TEA data on enrollment for the 2000-2001 academic school year).

1. Survey Objectives

The research goals of the Baseline survey are to establish a panel of sophomore and senior high school students in the state of Texas that can be followed to examine the decision making, knowledge and attitudes of students regarding post high school life course decisions in light of the existence of the Top 10 legislation in Texas. The baseline survey was intended to establish benchmark measures. Follow-up surveys with a subsample of the students will be used to track the evolution of student decision making about college attendance among those who attend college (full time or part time) immediately after high school graduation as well as those who decide to attend college one or more years after graduation.

The Baseline survey objectives called for the collection of 33,000 to 35,000 completed interviews with sophomores and seniors in Texas public high schools using a sample survey design. A probability sample of 100 high schools was desired. Interviews were to be conducted in class using self-administered surveys. This would require district and high school cooperation with the survey effort.

Analysis was desired at multiple levels of the education system -- students, schools and districts. Because of the multi-level nature of the analytic goals of the study, a census of sophomores and seniors was desired within the schools that were selected into the survey (to facilitate multi-level analyses). At the student level, analyses were desired separately by racial/ethnic subgroup: non-Hispanic whites; African Americans; Asians and Hispanics. Moreover, analyses of likely college goers and non-college goers were desired.

2. Sample Design

Frame. The Texas Education Agency (TEA) data base of all Texas public high schools was used to create the sampling frame for the selection of high schools. We started with the collection of all Texas public high schools as of fall 2000 (i.e., representing schools that operated in 1999). Each record of the TEA data base of high schools represented a "Texas public high school." We then eliminated "ineligible" schools and combined others in order to form the Primary Sampling Units (PSUs) that were employed in the first stage of selection.

Frame preparation involved a three-phase operation:

- identification of "eligible" schools (via deletion of ineligible schools)
- creation of Primary Sampling Units (PSUs)
- creation of sampling strata

We commenced by identifying and deleting ineligible schools from the sampling frame. Schools were deemed ineligible if they were either too small for the top 10 law to have an impact, or if the school operated under special, non-typical circumstances (e.g., charter schools, special education schools). The following school types were deemed ineligible:

- very small high schools (with senior enrollments of 10 or fewer);
- charter schools; and
- schools totally devoted to special education.

After eliminating ineligible schools, the sample frame contained 1,258 eligible public high schools, representing 210,000 seniors and 267,000 sophomores.

The next step of frame development involved formation of Primary Sampling Units. In order to create statistical efficiency in the baseline and follow-up studies, PSUs needed to be constructed so that they contained a minimum number of students. That minimum was determined to be 90 enrolled seniors. (Ninety represented the within-PSU sample size for the follow-up study. Anything less would necessitate special weighting.)

High schools whose senior enrollments fell below 90 were "linked" to similar high schools (in the same district or in a neighboring district) so that the resulting "cluster" exceeded the minimum size threshold. PSUs were formed by assembling clusters of between one and five high schools. The goal was to minimize the formation of clusters of schools due to the cost of collecting data from multiple schools when a single clustered PSU was selected. A total of 651 PSUs were formed, each containing between 1 and 5 schools.

Many schools were sufficiently large so that they constituted a single PSU (without need of clustering). Other schools needed only to be paired with one other school to achieve the minimum size threshold. And some PSUs (mostly in rural areas) contained three to five high schools. Ultimately, a total of 651 PSUs were formed for the sample selection process. Once PSUs were determined, they were assembled into sampling strata for application of the sampling process.

Stratification. Because the survey focused heavily on student decision making about college and more generally about post-high school life, it was important to stratify on variables that would be associated with these constructs. We incorporated the following variables into our principal stratification scheme:

- Metropolitan area status (MSA vs. non-MSA, as defined by the 2000 Census)
- Feeder school status (i.e., feeder to UT Austin or Texas A&M vs. non-feeder school)
- Racial/ethnic composition:
    - low concentration of White students (under 30% white students)
    - medium concentration of White students (between 30 and 59% white students), and
    - high concentration of White students (60% or more White students)
- PSU Type (a separate stratum for "small" PSUs with senior enrollments of 155 or less, as well as an indicator distinguishing "multiple" schools per PSU versus "single" PSUs –i.e., those comprised of a single school).

Eleven principal strata were formed using specific combinations of these stratification variables. The resulting strata and their definitions are presented in Table 1.

Stratification was also effected by sorting the PSUs prior to their selection using a systematic sampling scheme. Implicit stratification employed (1) number of schools in a PSU cluster (ranging in values from one to five), and (2) geography (i.e., region of the state -- Northeast; Northwest; Southeast; Southwest – as defined by the Texas Education Agency). PSU cluster size essentially served to stratify by school enrollment (i.e., size of school).

Table 1
Eleven Principal Strata Used in the Texas Higher Education Opportunity Project Baseline Sample of Public High Schools

Stratum Number:	PSU Type	Schools per PSU	Metropolitan Status	Ethnicity (%Non-Hispanic Whites)**
5	Small PSUs*	Multiple	N/A	N/A
21	Feeder PSU	N/A	N/A	HIGH density
22	Feeder PSU	N/A	N/A	MEDIUM density
311	N/A	Single	Metro Area	HIGH density
312	N/A	Single	Metro Area	MEDIUMdensity
313	N/A	Single	Metro Area	LOW density
321	N/A	Multiple	Metro Area	HIGHdensity
322	N/A	Multiple	Metro Area	MEDIUM density
323	N/A	Multiple	Metro Area	LOW density
421	N/A	N/A	Non-Metro Area	HIGH density
422	N/A	N/A	Non-Metro Area	MEDIUM-LOW density

*** The "Small PSU" stratum is composed of schools with fewer than 155 seniors.**

**** The Ethnicity strata are defined as follows:**

- **HIGH density = over 60% White total (grade 9-12) student enrollment**
- **MEDIUM density = 30-59% White enrollment**
- **LOW density = under 30% White**

Two stage probability sample. The sample design for the Baseline survey involved a two-stage probability sample of public high school seniors and sophomores in the State of Texas. At the first stage of selection, 62 PSUs were selected via stratified sampling with probabilities proportional to a measure of size equal to senior enrollment. Thus, the probability of selection for a school (*i*) is π, where:

$$\pi_i = (62)x\left(n_i / \sum_j n_j \right)$$

Technically, independent samples were selected within each stratum. The sampling method used to determine the number of PSUs per stratum leads to the selection probability that appears above.

The sample of 62 PSUs resulted in a total of 108 distinct high schools.

At the second stage of selection, all seniors and all sophomores were selected for the survey. In essence, a census of sophomores and seniors was taken within high schools. Since the sampling fraction at the second stage of selection is 1.0, the overall selection probability of a student (λ) in a high school is simply equal to the high school selection probability (π):

For student (*k*) in high school (*i*):

$$\lambda_{ik} = \pi_i \times 1.0 = \pi_i$$

Thus, schools are selected with unequal probabilities, and because we seek a census of senior and sophomore students within schools, then students come into the sample with unequal probabilities as well. This means that analytic weights are needed if unbiased estimates of population parameters are desired. The methodology used to generate analytic weights is described later in this report.

3. Survey Instruments

In this section we discuss the survey instrument and the pretesting process. The survey instruments were drawn principally from the senior and sophomore questionnaires employed in the 2001 Pilot Study. The survey instruments were converted to scannable documents that required students to use the bubble response format (like that used in standardized tests). Rather than use separate question booklets and answer sheets, an integrated question-and-answer format was preferred in order to reduce the risk of students recording a correct answer in an incorrect location.

Separate questionnaires were developed for seniors and sophomores. The senior and sophomore versions contained a common core, but also had measures appropriate for each respective grade level. The senior instrument was 24 pages in length, while the

sophomore version was 14 pages long. The principal difference is that seniors are asked questions about college applications and college perceptions that are not applicable to the sophomore cohort.

Both questionnaires contained a front cover that presented an informed assent statement (noting the voluntary nature of the survey and pledging confidentiality) and requested that the subject assent to participate by signing the form in the indicated space. Both questionnaires contained a back cover that requested contact information to be used in a follow-up survey.

Questionnaires were color coded so that the senior and sophomore versions could be easily differentiated. Questionnaires were uniquely identified using a sequential five-digit numbering scheme that appeared on each page of the questionnaire. These digits became the respondent identification number.

Questionnaires were printed on perforated paper so that all identifying information could be immediately separated from the subjects' responses, with only the ID number linking them to the subject's identity.

The instrument was pretested with 10 sophomores and 10 seniors. Students were recruited from several Austin area high schools (not of which were selected into the survey) and represented a diversity of socio-economic and racial ethnic backgrounds. Students came to a centralized location and were asked to complete the questionnaire and then participate in a debriefing session. Participants received a $10 incentive for their participation. The debriefing included the use of cognitive testing questions to explore comprehension and response construction process.

The questionnaire formats needed to be precisely arranged in order for the scanning software to be used effectively. The final arrangement of response bubbles was made by scanning software experts in partnership with the printers. Approximately 50,000 questionnaires were printed.

4. Field Methods

The project's approach to fielding the study involved a four-phase protocol:

1. contacting the districts of the sample schools to secure permission to contact the school;
2. after securing district approval, contacting the school principal to secure participation in the survey;
3. arranging dates and times for field staff to visit the school for in class survey administration;
4. implementing a mail survey for some of the schools that refused in-class participation but permitted mail surveys.

District cooperation. Securing cooperation from districts typically involved a three-step protocol:

- issuing an advance notification of the study to the superintendent;

- making telephone contact of the superintendent by Principal Investigator or co-Investigator to discuss the study and request permission to contact the school principal
- as requested, the preparing and submitting an IRB proposal for review and approval by the school district.

We note that in our IRB protocols explicitly requested use of parental passive consent and student written assent, even when district policy called for written parental consent. In all but two instances, district IRBs approved the project's use of passive parental consent.

High school cooperation. With the superintendent's permission, the school principal was contacted to request the school's participation. Securing school-level cooperation typically involved a two-step protocol:

- issuing an informational packet (via overnight courier) about the study
- contacting the principal by telephone to secure cooperation.

Logistical arrangements. Upon receiving an agreement to participate in the study, the project staff secured a point-of-contact at the school and began preparations for conducting the survey. The designated school contacts varied from school to school, and included the principal, the vice principal, registrar, counselors, technology coordinators, teachers, and administrative assistants.

Project staff elicited information about the specific school's *academic calendar* (e.g., the dates of TAKS testing, the week of spring break) and the *school class schedule* (e.g., use of A/B block scheduling where half of students' classes are covered across two days using longer class periods) because that determined the number of days that the survey would need to be conducted to capture all sophomores and seniors. This information was used to schedule the days required for conducting the survey.

To ensure that all seniors and sophomores were surveyed, we identified a core curriculum course -- English -- and requested that the school prepare a list of all English classes containing seniors and sophomores. English was used in virtually all schools because the Texas Education Agency requires high school students to take four years of English in order to graduate. And because of this requirement, it is not possible to "advance out" of grade-specific English classes (although there is a range of grade specific English classes – regular, enhanced, honors/pre-AP, and AP).

High school Registrars provided the project staff comprehensive lists of classes by, building, room number, teacher, class description (e.g., English I, II, III, IV), and counts of senior and sophomore enrollees in each class. Our communications protocols made it very clear to Registrars that we sought ***all*** sophomores and seniors, regardless of their matriculation through he English course series.

The lists were used to prepare teacher/class specific packets of parental consent forms. These were shipped to the point-of-contact two weeks before the scheduled survey date, along with blank copies of the questionnaires for the schools to have on file for parents and school staff. The packets were distributed to teachers, who in turn distributed the consent forms to students with instructions to give the forms to their parents. Students then would return to their teachers the signed nonparticipation directives if their parents

did not want the child to participate in the study. For this study the number of parental refusals was negligible, averaging about 5 refusals per school.

Survey Administration. The day before the scheduled survey date, project staff traveled to the high school to meet with the principal and the school administration, and to set up a staging area for the training of teachers and the distribution of survey materials. Typically, two project staff traveled to each school; for the smallest schools, only one staff person was needed.

A *survey operations center* was established in each school. It was typically housed in the school library. On the morning of survey administration, project staff arrived one to two hours before the start of classes. Tailored teacher-class survey packets were arranged alphabetically by teacher name and class level to facilitate the distribution of survey materials.

Teachers were instructed to report to the staging area a half hour prior to the commencement of classes. The project staff then conducted a training session to instruct teachers on in-class survey protocols. Training covered:

- purpose of the study
- voluntary nature of the study (e.g., withhold surveys from students whose parents had refused)
- confidentiality protocols (e.g., store completed surveys in sealed envelope)
- briefing material for the students (i.e., what to tell them)
- in-class survey distribution and collection protocols
- how to handle questions.

Teachers were then given their survey packets, after which they proceeded to class. A project staff person was available to introduce the study to classes, handle questions and answers, and monitor the data collection activity. Students were given 30-45 minutes to complete the questionnaire. Depending school-specific class schedules, this could involve the entire class period, or most of the class period. When a portion of the class was to be used for survey administration, teachers were asked to commence the survey administration with 30 minutes remaining in the class. That way the remainder of the class period could be devoted to the in-class survey. Although it was not possible to time each student, in-class observations suggest that most sophomores completed the survey in 25 minutes, while most seniors required 30 minutes (or more whenever possible).

At the end of each class period, completed surveys were placed in a large envelope (for only that class). Project staff then took possession of the packets and stored them securely in sealed, secure boxes at the school survey operations center. This box was under the control of the project staff at all times. Throughout the day, as each class period ended, survey materials were sealed, retrieved, and securely stored. When survey administration spanned more than one day, only the specific day's survey packets were issued (to avoid unnecessary loss or misplacement of materials).

Exceptions. All survey field plans encounter unanticipated problems, and the Texas Higher Education Opportunity Project was no exception. The following list documents the challenges encountered during the field period:

- Two high schools insisted that written parental consent be adopted. Just over a third (327/869 = 0.38) of the students from these schools returned written parental permission forms and were surveyed. Fortunately, one of the schools had a small enrollment (32 seniors and 51 sophomores) and was linked (to form a PSU) with a larger participating school that used passive consent. So the impact of that school was negligible. The other school was large (786 seniors and sophomores). We secured 318 surveys from this school using two data collection trips. The first trip was met with resistance by several teachers, so a second trip was required to conduct surveys in those classes where teachers were initially recalcitrant.
- Two other high schools cooperated only if the burden of the survey was reduced by subsampling classes. Moreover, students were required to complete their surveys outside of class time. To address these concerns, two project staff visited the school, systematically sampled half of all English classes from a list universe list (sorted by cohort) prepared by the Registrar, and had the teachers in the selected classes distribute surveys to students. Students were instructed to complete the survey on their own time and return it to project staff in the survey operations center located at the school library. Project staff stayed at the school for two consecutive days and took receipt of completed questionnaires. Students who returned a completed the questionnaire were offered a $5 cash incentive for their participation. Students returning completed forms were required to list their name and sign a log confirming receipt of the incentive payment. Only one staff person was allowed to handle the receipt process to reduce the risk of multiple entries. Students were also offered a postage paid return envelope if they were unable to complete the survey immediately but wanted to participate. This strategy proved highly successful. Just under three out of four questionnaires that were distributed were completed and returned. Although contamination (i.e., juniors, freshman and persons not from that high school) is a possibility, it was unlikely given the quick turnaround required for this procedure and the few mail returns that were received. It is not possible to determine which if any cases were ineligible.
- Twelve (additional) schools initially refused to participate but later agreed to a mail survey. Probability samples of 400 seniors and 400 sophomores were drawn from student directories. Samples were drawn independently for each school using stratified systematic sampling. Within a school the principal stratification variable was cohort (sophomore, seniors). Implicit stratification was effected using systematic sampling from the directory list sorted by last name. This helped to control the selection of students in a given cohort from the same family (since they typically would share the same last name). Our mail survey protocols involved an initial survey packet issued via priority mail to the parents of the student (requesting that parents review the survey and give it to their child to complete). The packet included a $5 advance incentive. A post card follow-up was sent two weeks later, followed by a final mailing of a survey packet to nonresponders. Overall, a 40% response rate was achieved for students selected into the mail survey component of the study.

- One high school refused to participate because the principal took exception to a number of survey questions about the student living arrangements and demographics. The school's participation was secured when project staff agreed to excise the problematic questionnaires from the survey instruments (by masking them so that the could not be seen). This affected 174 participants from a school that required mailing.

- One (additional) high school participated using in-class surveys but failed to remember that a portion of the seniors would be absent due to an extended trip; the traveling seniors were identified and surveyed by mail.

Field periods. Solicitation of district and school participation commenced in January 2002. The solicitation process ended in May 2002.

In-class survey data collection spanned a fourteen-week field period, commencing March 4, 2002 and concluding May 27, 2002.

The mailing survey process began in early May. Although returns were expected through June, questionnaires were received in July, August and even September. Acceptance of questionnaires was halted after a two-week period in which no surveys were received.

5. Results

School Participation. A total of 86 high schools participated in the survey. This represents a 93.3% response rate as shown in Table 2. Eight out of nine cooperating schools allowed in-class survey administration. And just under 2 percent of eligible schools refused to cooperate. Finally, 3 out of the 108 were found to be ineligible to participate in the study (i.e., exclusively servicing special education students).

Table 2
Disposition of Schools in the Texas Higher Education Opportunity Project

Final school participation status:	N of Schools	Percentage	Normed %	Response Status
In-class survey conducted	86	79.6%	81.9%	participated
Mail survey conducted	12	11.1%	11.4%	participated
Unable to secure cooperation	5	4.6%	4.8%	nonresponse
District/school refusal	2	1.9%	1.9%	nonresponse
Out of scope (ineligible)	3	2.8%	na	na
Total Number of Schools	**108**	**100%**		
			School Level Response Rate:	**93.3%**

With regard to Primary Sampling Units, schools in all 62 PSUs were represented by at least one constituent participating school. That is, all nonresponding schools were associated with PSUs comprised of school clusters that contained one or more participating constituent schools. This means that at the PSU level, the study achieved 100% representation.

Mail survey results. The results of the mail survey portion of the Baseline survey appear in Table 3. A total of 1,690 out of 4,200 mailed surveys were returned. This produced a 40.2% response rate for this component of the survey. The response rate was considerably higher among sophomores compared to seniors.

Table 3
Texas Higher Education Opportunity Project Baseline Survey
Student Level Response Rates for the Mail Component

Class:	Completes	Response
Seniors	696	34.8%
Sophomores	994	45.2%
Overall	**1,690**	**40.2%**

In-class survey results. The results of the in-class survey portion of the Baseline survey appear in Table 4. A total of 32,082 in-class surveys were collected. Sophomores (at 81% response) cooperated at a rate roughly 11 percentage points higher than seniors (at 70% response). This produced a 76.2% student level response rate for the in-class component of the survey. Like the mail component, the response rate was considerably higher among sophomores compared to seniors. PSU-specific response rates appear in Appendix A at the end of this report.

Table 4
Texas Higher Education Opportunity Project Baseline Survey
In-Class Student Level Response Rates

Class:	Completes	Response
Seniors	13,107	70.3%
Sophomores	18,975	80.9%
Overall	**32,082**	**76.2%**

Overall Results. A total of 33,772 completed questionnaires were gathered and processed. The distribution of cases by class is presented in Table 5. The overall

student level response rate -- obtained by combining mail and in-class components -- was 73.0%.

Table 5
Texas Higher Education Opportunity Project Baseline Survey
Overall Student Level Response Rates

Class:	Overall Completes	Response
Seniors	13,803	66.9%
Sophomores	19,969	77.9%
Overall	**33,772**	**73.0%**

5. Data Processing

Surveys forms were comprised of booklets with perforated pages that could be burst to remove identifying information from the survey responses. The data processing operation involved:

- receipt control of the questionnaires (i.e., logging returns)
- bursting the questionnaires to separate the identifying information from survey responses
- storing the contact and cover sheets in a separate, secure area from that of the answer sheets
- scanning the survey sheets
- conducting verification of the scanned responses
- coding of the college and country text responses
- conducting data cleaning
- developing analysis weights
- appending additional administrative data to the survey data set
- archiving the survey images.

Receipt Control. Once project staff returned to the central office from a survey field trip, their collection of completed questionnaires was reviewed, sorted, counted, logged and prepared for scanning.

Bursting. Project staff burst the questionnaire booklets to separate the student's identifying information from their survey responses. This was performed in numbered batches of questionnaires representing the questionnaires from part of one school (if large) or from several small schools. Creating these batches facilitated the subsequent scanning job. Batches of completed, burst survey forms were then sent to project staff who were responsible for the scanning operation.

Storing. The contact information sheets and cover pages were stored separately from the completed survey forms. They were stored in a locked file cabinet in a locked room. The survey forms themselves contained virtually no identifying information, since neither the high school nor the participant were identified anywhere on the answer sheets.

Essentially, the survey sheets are anonymous so long as access to the contact and cover sheets does not occur. Nonetheless, the survey response sheets were kept in a locked work room throughout the scanning process, as well as afterwards.

Scanning. The survey sheets were scanned using high speed scanning equipment. Use of the bubble form response format ensures a high degree of accuracy in the scanning operation (higher than survey industry standards). The survey answers were scanned first. After this was completed, the contact information was scanned as a separate scanning operation.

Verification. Because human action is required to place a response onto the survey form, there are inevitable instances when the scanning equipment cannot discern which is the appropriate response, or if there even exists a response (e.g., the subject lightly marked a response, marked outside the indicated bubble, erased, recorded multiple responses, etc.). In such cases the scanning software automatically flags the affected item and stores the image for a verification clerk to retrieve, inspect the markings and determine what if any response should be recorded.

Coding. A limited amount of coding was performed on the survey responses, limited to the respondents' preferred colleges/schools, and country of birth. For these items, the text response images were examined and Windows based (point and click) look-up tables were used to identify the IPEDS code of schools, as well as countries of origin. Coding protocols called for a minimum of 10% verification, plus checking of 100% of unusual collections of cases (e.g., anomalous or rare entries).

Data cleaning. Restrictive data cleaning was performed on the survey data. The principle we followed was to preserve the subjects' responses regardless of whether or not they were logical. That way the Investigators could determine from a substantive perspective the appropriate way to handle the data in analysis.

Null responses (i.e., no response recorded to a question) were processed to determine true "missing" or "inapplicable" status. "Missing" status means that the subject was supposed to answer the question but did not (for whatever reason). "Inapplicable" status denotes those instances when a question was properly left blank by following a skip instruction in the questionnaire.

There was one situation where a "missing" response could be changed to a valid response -- it involved deductive imputation. If a response could be deduced with certainty from the response of a subsequent question, then deductive imputation was allowed. For instance, if a subject left blank a question asking about college preference yet named a university in the follow-up "specify" question, then a "yes" could be deduced to the original question asking about the existence of a preference.

Appending Administrative Data. Some administrative data from the sampling frame and (sample management) control file were added to the Baseline survey data file. This included:

- TEA School Identification code
- Magnet School status
- School name
- District name

- Consent flag (passive vs. written parental consent)
- Data collection mode (mail, in-class, subsampling, cash incentive)
- Sampling strata (used in the sample design)

Archiving survey images. The images of the hard copy survey response pages were stored digitally for use in quality assurance and for posterity. The images are stored on CD. The data themselves are largely anonymous, identified only by sample number. The contact sheets and assent forms were not scanned but are stored as hard copy and were entered electronically via direct data entry.

Analysis weights. Analytic weights were developed for the survey data. The weights reflect two components:

- a sampling weight
- post-stratification adjustment that corrects for student level and school level nonresponse.

The sampling weight is simply the reciprocal of the probability of selection. Excepting those schools that required subsampling, the selection probability of a student is equal to the selection probability of the school. Here, the probability of a student (λ) in a high school is simply equal to the high school selection probability (π):

$$\lambda_{ik} = \pi_i \times 1.0 = \pi_i$$

And for those schools where a subsample of students was drawn from student directories, the selection probability of a student is simply the product of the school selection probability and the probability of selection of the student from the student directory. The probability of a student (λ) in a high school is simply equal to:

$$\lambda_{ik} = \pi_i \times \omega_{i\,j}$$

where π denotes the school selection probability and ω is the probability of selection student (*j)* from school directory (*i*).

Thus, sampling weight is simply the inverse of the λ for each student (*k*):

$$w_k = 1/\lambda_k$$

The post-stratification adjustment aligns the school level student totals to published TEA enrollment figures within our sampling strata (see Table 1 for a description of the strata). For each of the 12 sampling strata (*h*), known totals of senior and sophomore students were calculated using the TEA master frame used to draw the sample.

The post-stratification adjustment, A(*h*), was developed by taking ratio of *published* population totals T(*h*) for each PSU to the *weighted survey totals* S(*h*) for each PSU *k*:

$$A(k) = T(k)/S(k)$$

where

$$T(k) = \sum_i t_{ik}$$

(*i* denotes summation across all schools in the TEA school file that belong to PSU *k*) and

$$S(k) = \sum_i w_i \delta_{ik}$$

(δ denotes an indicator variable that is equal to 1 if a student is in a school belonging to PSU *k* and equals 0 otherwise; the *i* denotes summation across all students in the survey file whose schools belong to PSU *k*).

The *final analytic weight* θ is the product of the selection probability and the post-stratification weights:

For each student *i.*

$$\theta_i = w_i \times A_i$$

The PSU-level adjustments incorporate student level nonresponse as well as the nonresponse that occurred at the school level. Moreover, because adjustments are made at the PSU level, the results ensure that the weighted data will align with the published totals across sampling strata.

Sampling errors. Because the sample design employed a two-stage unequal probability sample design, the estimates of statistical precision provided by statistical packages that presume a simple random sample cannot be employed. The complex nature of the sample must be taken into account. To do this we employed a sampling error variance program called WestVar, Version 4.0.

Sampling errors were calculated using a Jackknife replication method (See, Wolter, K. Introduction to Variance Estimation, Chapter 4, Springer-Verlag: New York, 1985). For both the Senior and Sophomore samples, the sampling errors were calculated using weighted data; was based on the stratification used in the sample design.

Sampling errors were calculated for categorical and continuous variables. For categorical variables, sampling errors were calculated for the proportion of the sample in each category. For continuous variables, sampling errors were calculated for the arithmetic mean. Missing data were omitted.

Sampling errors were calculated for the entire sample and for three subclasses: Blacks, Hispanics, and Other. Average Design Effects were calculated and appear as Appendix B, Tables B1 and B2 at the end of this report.

The square roots of the average DEFFs – referred to as “DEFTs” are often used to correct simple-random-sample-based (unweighted) sample sizes to generate adjusted sampling errors. These adjusted sampling errors will reflect the actual, complex sample design.

To illustrate, if 12.5% of Hispanic seniors are following a Distinguished Achievement graduation plan (Q3 in Section A), and their unweighted sample size for seniors is 3,538, then under a simple random sample, the variance of this estimate would be:

SRS Var(p) = p(1-p)/n = (0.125)(.0.872)/3538 = 0.00003085

The square root of this value is used for confidence intervals and test statistics:

SRS Sampling Error (p) = SQRT(0.00003085) = 0.00555

This value understates the actual precision of the sample. To incorporate the impact of the complex sample, we consult Table B3 (DEFTs) for Seniors. The average DEFT for questions in Section A for Hispanics is 1.92. The adjusted sampling error would be:

Actual Sampling Error (p) = SQRT{ [p(1-p)/n] }*DEFT

= (0.00555)*1.92 = 0.0107

To illustrate the use of the adjusted sampling error, suppose we desire a 95% confidence interval for the percentage of Hispanic seniors who follow a distinguished program. The calculations would be:

95% Confidence Interval: 12.5% +/- (1.96)* (1.07%) = 12.5% +/- 2.1%.

APPENDIX A
Average Design Effects for the Texas Higher Education Opportunity Project Baseline Survey

Table A1: Average Design Effects for Seniors by Race-Ethnic Group and Instrument Section

Senior DEFFs	WHITE	BLACK	HISPANIC	Overall
Section A: Course Taking	5.74	2.49	3.67	6.68
Section B: Test Scores/Guidance	4.66	2.63	3.35	5.46
Section C: College Admission Knowledge	4.36	2.11	2.71	4.35
Section D: Future Plans	4.44	2.32	2.75	4.81
Section E: College Perceptions	4.00	1.96	2.71	4.50
Section F: Demographics	3.76	2.31	3.84	5.50

Table A2: Average Design Effects for Seniors by Race-Ethnic Group and Instrument Section

Sophomore DEFFs	WHITE	BLACK	HISPANIC	Overall
Section A: Course Taking	6.03	2.89	4.75	7.47
Section B: Test Scores/Guidance	15.28	3.39	5.86	13.28
Section C: College Admissions Knowledge	4.20	2.04	2.84	6.87
Section D: Future Plans	3.21	1.82	2.57	5.27
Section E: Demographics	8.87	2.41	4.68	11.17

Table A3: Square Roots of Average DEFFs for SENIORS by Race-Ethnic Group and Instrument Section

Senior DEFTs	WHITE	BLACK	HISPANIC	Overall
Section A: Course Taking	2.40	1.58	1.92	2.58
Section B: Test Scores/Guidance	2.16	1.62	1.83	2.34
Section C: College Admission Knowledge	2.09	1.45	1.65	2.09
Section D: Future Plans	2.11	1.52	1.66	2.19
Section E: College Perceptions	2.00	1.40	1.64	2.12
Section F: Demographics	1.94	1.52	1.96	2.35

Table A4: Square Roots of Average DEFFs for SOPHOMORES by Race-Ethnic Group and Instrument Section

Sophomore DEFTs	WHITE	BLACK	HISPANIC	Overall
Section A: Course Taking	2.46	1.70	2.18	2.73
Section B: Test Scores/Guidance	3.91	1.84	2.42	3.64
Section C: College Admission Knowledge	2.05	1.43	1.68	2.62
Section D: Future Plans	1.79	1.35	1.60	2.30
Section E: Demographics	2.98	1.55	2.16	3.34