

# Chapter 13: Sampling

## Overview

Sampling plays a critical role in the research and annual development activities necessary to support the Texas Student Assessment Program. Through the careful selection of student samples, the Texas Education Agency (TEA) is able to gather reliable information about student performance on its assessments while minimizing campus and district participation. Because the results from a well-drawn sample can be generalized to the overall student population, the way in which a sample of students is selected is critical. This chapter discusses the key concepts of sampling, the reasons for sampling, types of sampling designs, field-test sampling, sampling designs for research studies, and the role of sampling in the assessment program.

## Key Concepts of Sampling

### Target Population

A target population is the complete collection of objects (for example, students) we want to learn something about (Lohr, 1999). This is the set of students for whom we want to be able to generalize the results. For example, for a study with the goal of understanding how grade 3 English language learners (ELLs) perform on a set of test questions, the target population may be all grade 3 ELL students in Texas. Defining the target population is an important but difficult task in sampling. Therefore, careful consideration is given to defining the target population prior to sampling.

### Sampling, Sample and Observation Unit

Sampling is the process of selecting a subset of the target population that will allow reliable and valid inferences about the target population. The primary goal of sampling is to create a small group from the population that is as similar as possible to the larger population. A sample is a subset of the target population that will participate in the study.

A sampling unit is the unit to be sampled from the target population. A sampling unit could be a student, a campus, a district, or even a region. For instance, if a study selects all students on 20 campuses from a list of all campuses in the state, then campus is the sampling unit of this design.

An observation unit is the unit on which data are actually collected. An observation unit could be the same as the sampling unit. For example, each of the 20 sampled campuses could report the number of computers it has. In this case, the observation unit is also campus. An observation unit could also be smaller than the sampling unit. If data are

collected on all the students in the selected campuses, then student is the observation unit.

## Reasons for Sampling

Texas employs sampling instead of studying entire target populations for several reasons, including:

- **Size.** It is more efficient to examine a representative sample when the size of the target population is quite large.
- **Accessibility.** There are situations where collecting data from everyone that forms the target population is not feasible.
- **Cost.** It is less costly to obtain data for a selected subset of a population than it is for the entire population.
- **Time.** Using the tool of sampling to study the target population is less time-consuming. This consideration may be vital if the speed of the analysis is important.
- **Burden.** Sampling minimizes the requirements for campus and district participation in field testing and research studies, reducing testing burden.

## Types of Sampling Designs

### Probability Sampling

In a probability sample, all sampling units have a known probability of being selected. This means that the number of sampling units in the target population is known and can be listed. For example, if student is the sampling unit, then we need to know the number of students in the target population and have a list of all the students. Random selection from the list of sampling units is a key component in probability sampling. The major probability sampling designs include: simple random sampling, stratified sampling, and cluster sampling. Each of these is described next.

### Simple Random Sampling (SRS)

In simple random sampling (SRS), sampling units are randomly selected from the list of all sampling units. Random selection means that all sampling units in the target population have the same probability of being selected. For example, an SRS of grade 3 students is constructed by randomly selecting from the complete list of all grade 3 students in Texas with each student having the same chance of being in the sample. Simplicity in making inferences is one of the main advantages in SRS. Another advantage is that every sampling unit (e.g., student or campus) has an equal chance of participation. However, SRS can result in a nonrepresentative sample of certain characteristics, such as ethnicity, gender, socioeconomic status, or geographical location. For example, because all students in the state have an equal chance of being selected in an SRS, it is possible

(though unlikely) to have a disproportionately high percentage of students from North Texas.

## **Stratified Sampling**

Given the same number of sampling units, stratified sampling often provides a more representative sample than does simple random sampling. Under this design the list of sampling units is first grouped (or stratified) based on certain characteristics. A simple random sample is then taken for each group (or stratum). For example, all Texas students can be grouped by region; then students are sampled randomly from each region. This way, a known percentage of sampling units with characteristics based on the grouping variables such as geographical region, gender, or ethnicity, are always in the sample. Consequently, stratified sampling typically produces a more representative sample and leads to more accurate estimation of the parameters of interest.

## **Cluster Sampling**

Another common probability sampling design is cluster sampling. With cluster sampling, the list of all sampling units is first grouped into clusters based on certain characteristics or variables of interest. Then, unlike stratified sampling, a predetermined number of clusters are selected and all sampling units within the chosen clusters are observed. For example, all Texas campuses can be grouped into regions. Then, a predetermined number of regions can be selected and all campuses within the chosen regions would be selected.

## **Nonprobability (Convenience) Sampling**

A sample that is created without the use of random selection is called a nonprobability (or convenience) sample. In convenience sampling, there is no listing of complete sampling units, and sampling units have no known probability of being selected. The sample may consist of volunteers or students from local campuses. Many research studies conducted at universities involve nonprobability samples of college students because of the ease of obtaining participants. The sampling is literally conducted for convenience, which introduces sources of potential bias into the resulting data.

Sampling and probability theories do not readily apply to studies conducted with convenience sampling. Additionally, convenience sampling makes it difficult to generalize results to the target population. However, in certain situations, convenience sampling is conducted because it is the only viable option. A list of students or campuses in the target population may not exist. For example, early studies involving the Texas Assessment of Knowledge and Skills-Alternate (TAKS–Alt) assessment have used convenience samples. This is because a list of students for whom TAKS–Alt is the appropriate assessment was not available prior to the first operational administration. In other cases, a random sample may not be necessary to accomplish the goals of the study. This may occur when a study is small and the first in a series of studies planned to address the same question. In most

cases where convenience sampling is conducted, the representativeness of the sample needs to be considered and mentioned when drawing conclusions.

## **Sampling with or without Replacement**

Regardless of the type of probability sampling design used, one decision that needs to be made is whether to sample with or without replacement. To help illustrate the distinction between the two sampling methods, consider SRS with or without replacement.

Suppose a simple random sampling with replacement (SRSWR) of size  $n$  is obtained from a population of  $N$ . This means that each time a sampling unit is randomly chosen, it is placed back into the target population and could be chosen again. In other words, with SRSWR, it is possible for any given sampling unit to be selected multiple times and have its data duplicated in the resulting sample of size  $n$ .

On the other hand, in a simple random sampling without replacement (SRS) of size  $n$  from a population of  $N$ , each chosen sampling unit is ineligible to be selected again in the sample. Thus, under SRS, each sample consists of  $n$  distinct, nonduplicate units from the population of size  $N$ .

Typically, SRS is preferred over SRSWR because duplicate data adds no new information to the sample (Lohr, 1999). The method of sampling with replacement, however, is very important in resampling and replication methods, such as bootstrapping.

## **Resampling and Replication Methods: Bootstrap**

Resampling and replication methods, such as bootstrapping, treat the sample like a population. They repeatedly take pseudo-samples from samples to estimate variances and standard errors. Thus, sampling with replacement is assumed with these methods.

The bootstrap method was developed by Efron (1979), and described in Efron and Tibshirani (1993). Suppose  $S$  is an SRS of size  $n$  from the target population. Pseudo-samples, SRS samples of size  $n$ , are taken with replacement from  $S$ , and are assumed to be like samples of size  $n$  from the population. A pseudo-sample from  $S$ , however, is most likely not the same as  $S$  since the pseudo-sample is a resample with replacement. This process is repeated  $R$  times. The variance and standard error estimates are finally obtained, using the  $R$  sets of pseudo-samples. In Texas, comparability studies which compare online and paper versions of a test form have used the bootstrap method.

## **Sampling in the Texas Student Assessment Program**

Sampling is used in the assessment program when not all eligible students participate in an assessment activity. Two major types of these activities are: 1) testing as part of a research study, and 2) stand-alone field tests. Research studies in general involve assessing a subset of students, sampled and assigned to various testing conditions, from

the state population to support the reliability and validity of the assessment program. Field-test results are used to evaluate statistical properties of newly-developed items prior to their use on a live test form. While Texas employs embedded field testing wherever possible, there are situations in which stand-alone field testing is necessary. This may occur when test structure, small student populations, new tests, or method of test delivery preclude embedding field-test items. Samples for stand-alone field tests are selected to mirror important characteristics of the state population such as ethnicity and campus size. The results can then be generalized and utilized to make recommendations on item selections for use on future Texas assessments. In rare situations, census field testing can be conducted. For instance, the entire target population rather than selected samples participated in TAKS–M field tests in fall 2007 and spring 2008.

The remainder of this chapter will discuss the major methods used for field-test sampling for stand-alone field tests in 2007–2008 and will provide an overview of the methods used for sampling for research studies conducted in 2007–2008. Census field testing was used in this situation because the group of students taking TAKS–M is quite small (small student population) so all students were required to participate in the field test to make sure enough data were available to conduct field-test analyses.

## **Field-Test Sampling**

TEA recognizes the challenges districts and campuses face with regard to time and resources when they are asked to participate in stand-alone field tests. Changes in sample sizes and the sampling process were made during the 2007–2008 academic year to keep district field testing to a minimum while maintaining the high quality of the assessment program.

After discussion with the Texas Technical Advisory Committee (TTAC) in spring 2007, a decision was made to reduce the sample size for TAKS English and EOC stand-alone field tests forms from a minimum of 500 students per analyzed subgroup to a minimum of 280 students per analyzed subgroup. In general, the number of students and campuses sampled for TAKS in 2008 was reduced by approximately 50 percent from the number sampled in 2007. Although a similar number of campuses and districts participated in the EOC field tests in 2008, fewer students per campus participated. The maximum number of students participating in an EOC field test at any campus was reduced from 300 to 100. Moreover, no stand-alone field test was conducted for TELPAS reading because of a new method developed for embedding the Texas English Language Proficiency Assessment System (TELPAS) reading field-test items in operational assessments.

## **End-of-Course Field Tests**

In spring 2008, the chemistry and U.S. history EOC field tests were given for the first time. Though stand-alone field testing was required for the first year of these tests, embedded field testing will be used in the future. The EOC stand-alone field tests used a stratified sampling design where campus was the sampling unit, but student was the

observation unit. TEA initiated a sampling model in spring 2006 that provided a “relief year” to campuses such that each campus would have a minimum of one of every five years during which they would not administer the TAKS stand-alone field testing. The relief year offered in TAKS was extended to include the EOC field tests such that campuses that had participated in the TAKS stand-alone field testing for the previous four years were exempted from selection in both the 2008 TAKS stand-alone field tests and the 2008 EOC field tests.

Additionally, the sample size for EOC forms and the maximum number of students participating in an EOC field test at any campus were reduced. Because the EOC field tests were administered online, campus technology infrastructure was considered. After considering a variety of factors, each campus selected for EOC field testing in 2008 was asked to test in both subjects and to test half of their enrolled students, regardless of grade, up to a maximum of 100 students in the testing window. In comparison to a maximum number of 600 students per campus (300 for each assessment) required for 2007 EOC field testing, the maximum number of students per campus required for 2008 EOC field testing was reduced to 200 students.

The EOC field-test samples were selected prior to the selection of TAKS samples to get the requisite number of students for the chemistry and U.S. history field tests. Campuses were selected for 2008 EOC stand-alone field tests after removing those

- that were scheduled for a relief year in 2008;
- that were selected for National Assessment of Educational Progress (NAEP) testing;
- that were defined as juvenile justice alternative education program (JJAEP), disciplinary alternative education program (DAEP), or Texas Youth Correction (TYC) facility; and
- that did not have at least 15 students.

The field-test sample was selected from the resulting set of campuses. Because the campus was the unit for sampling, it was necessary to obtain the student enrollment from each campus in order to produce an estimate of student counts. The student counts for each campus were based on the number of students who enrolled in the chemistry course and/or U.S. history course in fall 2006.

The selection proceeded as follows:

1. All eligible campuses were divided into five even-sized strata based on campus size.
2. If a number of campuses of equal size appeared around the cut between strata, the placement in the upper or lower strata was done randomly.
3. Campuses were sorted randomly within each stratum.

4. One campus was randomly selected from each stratum into the sample in ascending and descending order of strata (e.g., 5-4-3-2-1-1-2-3-4-5-5-4-3-2-1-...). One campus was selected in one stratum first before moving to the next stratum.
5. The number of students in the sample was evaluated relative to the target number of students after the campus had been selected. The above step was repeated until the target number of students was reached.
6. A “fit” index was calculated for the resulting sample of campuses. This index indicated how well the selected campuses reflected the demographic breakdown of the state population.
7. The above steps (from dividing campuses into five strata to calculating the fit index) were repeated up to 1000 times. Any sample for which the fit index indicated that the sample was within 0.5 percent of the target demographic breakdown was reviewed by a psychometrician, who selected a final sample using professional judgment.
8. Once the final sample was determined, it was regenerated using the appropriate random number seed so that additional detailed output descriptive statistics for this sample could be generated.

The final sample was determined after evaluating four key elements: fit to statewide ethnic percentages, number of campuses, number of students, and distribution of campus size strata within the sample. A summary of the number of campuses and students selected for the 2008 EOC field tests is provided in Table 4.

## **TAKS Stand-Alone Field Tests**

Sampling for the TAKS stand-alone field tests was conducted at grades 4 (English version) and 7 writing, 9 reading, and 10 and 11 ELA English language arts (ELA). The “relief year” sampling model that was initiated in 2006 was maintained in 2007 such that campuses who had participated in the TAKS stand-alone field testing for the previous 4 years were exempted from selection in 2007.

In addition to reducing the minimum number of students needed for each analyzed subgroup from 500 students to 280 students, a decision was made by content experts to reduce the number of forms on which a set of items associated with a triplet are field tested (for TAKS grade 11 ELA, TAKS grade 10 ELA, and TAKS grade 9 reading) from three forms per triplet to two forms per triplet. Psychometric services combined the information on number of forms needed for the 2008 field test with the number of students needed per form to determine the total number of students to sample for each test.

Campuses were selected for 2008 TAKS stand-alone field tests after removing

- campuses that were scheduled for a relief year in 2008;

- campuses that were selected for NAEP testing;
- campuses that were defined as JJAEP, DAEP, or TYC;
- campuses that did not have at least 15 students; and
- campuses that were selected for three other 2008 TAKS stand-alone field tests.

Similar to the sampling process for EOC stand-alone field tests, the student enrollment information from each campus was obtained to produce an estimate of student counts. Note that the TAKS grade 11 was sampled from within the selected EOC sample. In other words, all campuses selected to participate in the TAKS grade 11 stand-alone field tests also were selected to participate in the chemistry and U.S. history EOC field tests. The specific procedure used to select the campuses in the field-test sample is outlined below.

- 1) The targeted number of students (using campus as selection unit) is selected from the grade 11 campuses that were already selected for EOC field testing.
- 2) To select students for grade 10 field testing, the targeted number of students was selected from the grade 10 campuses that were already selected for grade 11 field testing and the grade 10 campuses that did not have grade 11 students.
- 3) To select students for grade 9 field testing, the targeted number of students was selected from grade 9 campuses that were already selected for grade 11 field testing, and the grade 9 campuses that did not have grade 11 students.
- 4) Campuses that had been selected for field testing at ALL three TAKS grade levels (9/10/11) were excluded from further sampling. These campuses had reached the maximum number of grade levels allowed and should not have been selected to participate at other grade levels. After excluding these campuses, the remaining campuses were used to select grade 7 students.
- 5) Grade 7 students were selected from the remaining sample pool.
- 6) Campuses were excluded from the sample pool if they were selected for field testing at three of the four grade levels (grades 7, 9, 10, and 11). The remaining sample was used to select grade 4 students.
- 7) Grade 4 students were selected from the remaining sample pool.

Within each grade level, the selection procedures that involved the use of five strata were the same as described previously for 2008 EOC field tests. The final sample was also determined based on the series of indicators used in the EOC sampling process.

## Results

A summary of the number of students and campuses selected for the 2008 TAKS and EOC stand-alone field tests is provided in Table 4.

**Table 4. 2008 TAKS and EOC Stand-Alone Field Tests—Sample Summary**

	2008 Sample Number of Students	2008 Sample Number of Campuses	Number of Forms
Grade 4(E) Writing	46,387	581	20
Grade 7 Writing	44,752	247	17
Grade 9 Reading	39,992	180	14
Grade 10 ELA	58,149	293	20
Grade 11 ELA	68,797	361	22
EOC Chemistry	35,090	676	11
EOC U.S. History	37,937	676	11

**NOTE:** All samples were within 0.5% of the statewide ethnicity targets.

Additionally, the differences in number of campuses and students selected for the 2007 and 2008 TAKS and EOC stand-alone field-test samples are shown in Table 5.

**Table 5. Differences between 2007 and 2008 TAKS and EOC Stand-Alone Field-Test Samples**

	Difference in Number of Students	Difference in Number of Campuses
Grade 4(E) Writing	44,121	611
Grade 7 Writing	31,765	171
Grade 9 Reading	49,938	179
Grade 10 ELA	32,035	115
Grade 11 ELA	61,616	334
Total TAKS	219,475	1,072
EOC	75,993	19

**NOTE:** Differences reflect 2007 counts minus 2008 counts, so positive numbers reflect smaller samples in 2008.

In sum, the changes in the process for sampling for TAKS stand-alone field tests resulted in nearly 220,000 fewer students selected in 2008 relative to 2007. For EOC the reduction was about 76,000 students. This decrease represents 46 percent of the 2007 TAKS sample and 51% of 2007 EOC sample. In general, the number of students and number of campuses sampled for TAKS in 2008 was reduced by about 50 percent from the number sampled in 2007. The one exception to this was TAKS grade 10 ELA, where the reduction from 2007 to 2008 was about 30 percent. Although the number of forms per triplet was reduced between 2007 and 2008, the number of triplets field tested at grade 10 increased during that period, resulting in no net reduction in the number of forms. The percentage of usable data received for TAKS grade 10 ELA was lower than for other tests, so a larger number of students per form were sampled to account for this.

While the percentage of students sampled in 2008 for EOC was about 50 percent less than the number of students sampled in 2007, the difference in the number of campuses was very small. This reflected efforts to reduce the maximum number of students per campus required to participate in the field test from 300 to 100. Although a similar number of campuses and districts participated in the field test in 2008, fewer students per campus participated.

## Sampling Designs for Research Studies

### 2008 TELPAS Audits

The purpose of the TELPAS audits is to examine and monitor the efficacy of the TELPAS holistic rating process. Audits were performed for the domains of listening, speaking, and writing for grades 2–12. The listening and speaking audit was a pilot and was performed as a separate audit from writing.

The 2008 listening and speaking pilot involved two small-scale studies for grade 2 and grade cluster 6–8. Students in grade 2 are not tested as often as students in other grades, which lessens the statewide assessment burden for this grade. The selection of grade cluster 6–8 allowed for the evaluation of the logistics of following secondary students to different classes to rate their English proficiency in a variety of academic settings. The 2008 listening and speaking pilot utilized random sampling. Thirty students from each of the four proficiency levels (Beginning, Intermediate, Advanced, and Advanced High) within each grade cluster (grade 2, grades 6–8) were randomly sampled. As a result, approximately 150 students per grade cluster (a total of 309 students) participated in the audit.

The 2008 writing audit also utilized random sampling. As shown in Table 6, 125 students were randomly selected from each of the four proficiency levels across four grade clusters (grade 2, grades 3–5, grades 6–8, and grades 9–12). This resulted in a sample of 2,000 (125 students × 4 proficiency levels × 4 grade clusters) students in the 2008 writing audit.

**Table 6. Number of Students Sampled for 2008 TELPAS Writing Audit**

Grade Cluster	Beginning	Intermediate	Advanced	Advanced High	Total
2	125	125	125	125	500
3–5	125	125	125	125	500
6–8	125	125	125	125	500
9–12	125	125	125	125	500
<b>Total</b>	<b>500</b>	<b>500</b>	<b>500</b>	<b>500</b>	<b>2,000</b>