Quick Guide to Sample Sizes, Sampling & Representation For Program Assessment

Office of Assessment of Teaching and Learning, Washington State University, February 2019

Rather than assessing an entire group of students (a census), a sample assesses a subset of a particular population. A sample is often used when assessing an entire group of students is overly difficult or time consuming. This document provides tips for sample size and sampling strategies, along with examples for undergraduate degree programs. ***For assistance determining appropriate sample sizes and strategies for your assessment specific to your program context (including particularly large or small enrollments), or other questions, please contact ATL.***

Sampling for Assessment

There is no "one size fits all" approach to sampling for program assessment, but there are some ways to make it more approachable. For programs or courses that are small, assessing the entire group of students (a census) may yield a more accurate measure of student learning. On the other hand, a sample facilitates the assessment process when it is not feasible to assess all students—for example, when programs/courses have large numbers of students, when artifacts take a long time to evaluate, or when participation in an assessment is voluntary (e.g. a voluntary student focus group or survey).

It is important to situate your approach to sampling in the context of the type of assessment measure being collected (e.g. embedded assessment, survey, focus group, etc.) and the assessment questions that you are trying to answer. In other words, whether or not to sample and the size of the sample depend on multiple factors, such as:

- The number of students enrolled in the course or program, including any sub-categories of interest (e.g. major/option and campus)
- The length and complexity of the assessment measure/rubric/scoring tool and assignment/artifact
- If the assessment measure has been collected before or pilot tested
- The number of faculty members who will be involved in evaluating student work, as well as the number of faculty evaluating each artifact
- Factors beyond your control (e.g. the number of surveys or assignments completed, unintentional errors, etc.)

Examples of a Census:

- A program wants to examine student achievement in its capstone course. The course ends the semester with eight majors, each of whom is required to write a 10-15 page paper. All eight papers are evaluated by a group of faculty using a rubric to assess two program student learning outcomes (SLOs).
- A department gives a common online multiple choice exam to all 168 students across both sections of an introductory-level course. One of the program's SLOs is aligned with five of the questions on the exam. The scores for these questions are aggregated for both sections, comprising responses from all 168 students.
- A program offers a capstone course with a maximum enrollment of 25 students on each of three campuses. Each instructor assesses student posters using a program rubric for one SLO, for all of the students in their individual course. The scores are aggregated for all three campuses, comprising responses from all students in the capstone.

Examples of a Sample:

- In a program with roughly 60 seniors on each of two campuses, faculty assess student poster presentations for a random sample of 36 seniors on each campus using a rubric for two program SLOs.
- A department runs five sections of the capstone course involving 98 total students on a single campus. One program SLO is to be assessed during student oral presentations by the course instructor using a program rubric. Each instructor randomly selects 10 presentations from their course to evaluate, for a total of 50 students.
- Each year, a program administers an exit survey to all of its graduating seniors. Last year, the survey was completed by 78 seniors (out of the 158 total seniors).

Choosing a Sample Size

How many students need to respond to a survey for a program to be reasonably confident about the results of the survey? How many papers does a program need to collect to assess degree program learning outcomes? To answer these questions, consider the following.

Anytime you do not assess the entire group of students (a census), the results will have some margin of error. The level of error is measured as a percentage, as is the level of confidence. The level of confidence represents how confident you feel about your error level. For example, if you have a 90% confidence interval with an error level of 10%, you are saying that if you were to conduct the same survey 100 times, the results would be within +/- 10% of true population value 90 times out of 100.

When deciding an acceptable sample size, consider how much sampling error can be tolerated and what confidence level is acceptable. In other words, how much precision is needed? While this may change based on the types of decisions that the results from the assessment measure may guide, general recommendations are to select a margin of error no greater than 10% and a confidence interval of at least 90%. The following table can help you determine the level of sampling error and confidence interval associated with certain sample sizes. To calculate for other population sizes, see the calculator at http://www.custominsight.com/articles/random-sample-calculator.asp.

Completed Sample Sizes Needed for Various Population Sizes						
Population size	Sample size for a 90% confidence interval			Sample size for a 95% confidence interval		
	±15% Sampling Error	±10% Sampling Error	±5% Sampling Error	±15% Sampling Error	±10% Sampling Error	±5% Sampling Error
25	14	18	23	16	20	23
50	19	29	42	23	33	44
100	23	40	73	30	49	79
200	26	51	115	35	65	132
400	28	58	162	39	77	196

Note: The sample size calculations here pertain to clean, useable data from your assessment work. When planning, it is recommended that you include a few additional students or papers, so that you will be able to deal with incomplete data and unexpected situations (e.g., a student paper is missing pages, a rater skips a portion of the rubric, technology glitches, etc.).

General Feasibility Tips for Choosing a Sample Size: It is important to choose methods (and sample sizes) that are feasible given program resources and faculty time. These are intended to be general guidelines and are not hard and fast rules, as contexts vary greatly between programs (e.g. the number of students in a course or program, and the presence of any sub-categories of interest, such as major/option and campus).

- General Recommendations for Sample Sizes: If there are 40 or more students in the *population(s) of interest*, we suggest a representative sample of at least 40 students from each population of interest. If there are fewer than 40 students in the *population(s) of interest*, plan on collecting evidence from all students. In some cases it may be necessary to oversample from a particular group (see *Sampling Strategies*).
- If student work will be evaluated using a rubric by a faculty rater *other than the course instructor*, keep this in mind: in our experience, it takes a faculty rater at least 15 minutes to apply a rubric to score each short written project (such as a short essay, research poster, etc.) and even longer for more complex projects and rubrics. So if 6 faculty raters spend 90 minutes evaluating student work, that's (a sample of) roughly 36 students if each project is evaluated by only one faculty rater (or 18 students if scored by two raters).

Sampling Strategies

A sampling strategy is used to identify a subgroup that effectively represents the population as a whole. Below are four types of sampling: simple random sampling, stratified random sampling, self-selection sampling, and convenience sampling.

Simple Random Sampling

In a randomized sample, every student in the population (e.g., all seniors in your program) has an equal chance of being chosen to participate or having their paper selected for review. There are several ways to collect a random or semi-random sample. One method is to use a computer program (e.g. Excel or an online random number generator) that randomly selects respondents from the pool. A second method (also known as systemic sampling) is to select two random numbers; the first number tells where to start in a list of students or papers and the second random number indicates how many to count before selecting a second student for the sample. For example, if you chose 32 and 8, you would start with the 32nd student on a list and count down the list including every eighth student in the sample.

Stratified Random Sampling

Taking a stratified random sample involves dividing the population into sub-categories, and randomly selecting from each sub-category. A stratified random sample is taken when you want to ensure that the sample includes students from each group of interest (such as students from every option or campus). To stratify a population, you first need to decide what sub-categories are of interest and in which you suspect there may be substantial differences. Then, a simple random sample is selected from each group. Ideally, the percentages of students in each group in the sample will be the same as the percentages of each group in the overall population. For instance, if 25% of students are in Option A and 75% are in Option B, then the sample should include 25% students from Option A and 75% from Option B. In some cases it may be necessary to oversample from a particular group. For example, a program with 75 students on Campus A and 8 students on Campus B, may decide to include all 8 students from Campus B in the sample.

Self-Selection Sampling

Self-selection sampling allows participants to volunteer and/or decide if they would like to participate in an assessment. Self-selection sampling may be done by asking for volunteers (i.e. inviting students to participate in a voluntary focus group or survey). While every student may have an equal chance of being included in the sample (if all students are invited to participate), there is a potential for bias if certain groups in the population are less likely to participate (see *Considerations for Sample Representation*).

Convenience Sampling

Convenience sampling is often called grab sampling, and uses whatever participants from the population that are available to participate at a given time. This technique has very little structure and the only criterion for selection is that the participant you select is a member of the population and is available to participate at the time required. Convenience sampling does not use random sampling at any stage of the selection process, so some members of the population may have a greater chance of being selected. With convenience sampling, the potential for bias is high because the sample is made up of students that were simply convenient or available at the time.

For example, if you wanted to determine how satisfied students were with your degree program, it might be convenient to sample and survey every fifth student that came into the main department office. However, this method may only measure the satisfaction of students who choose to come into the office (see *Considerations for Sample Representation*).

Considerations for Sample Representation

In general, assessment data is collected locally to make local decisions. Since results do not need to be generalizable outside of a local context, the most important consideration is whether or not a sample is a representative of the entire local population. Samples are representative when they provide an accurate reflection of the variations and diversity represented within a population. For example, does the sample include both high-achieving and low-achieving students? Does the sample include a proportional number of students from all degree options?

A representative sample parallels the key variables and characteristics of the population, such as sex, age, campus, option, etc. In a classroom of 60 students, in which half the students are male and half are female, a representative sample might include 40 students: 20 males and 20 females.

Generally speaking, samples produced by random sampling will generate a sample that is representative of the entire population. However, when not every member of your target population has an equal chance of being included in the sample or some groups choose not to participate (such as with convenience or self-selection sampling), there is a risk that your sample may not be representative of the entire population. For example, the presence of non-response bias should be considered when evaluating the results of voluntary surveys, as there is concern that those who did not respond (non-respondents) may have different views than those who did respond and therefore the results are not representative of the entire group.

Many times the potential for bias is due to factors beyond your control (i.e. some students don't submit the assignment or respond to the survey, a student paper is missing pages, a rater skips a portion of the rubric, technology glitches, etc.). In these instances, it can be particularly useful to compare key variables and characteristics in your sample to those of the population. For example, the following table compares key characteristics for a sample of 110 seniors whose papers were assessed in capstone courses compared to the entire population of 224 seniors enrolled in capstone courses.

Sample Representation of Seniors in Capstone Courses		
Characteristic	% of students	
	Sample (110 seniors)	Population (224 seniors)
Option		
Option A	84%	84%
Option B	16%	16%
Sex		
Male	62%	65%
Female	38%	35%
End of Term GPA		
3.00-4.00	69%	67%
2.00-2.99	24%	25%
<2.00	7%	8%

Note: When examining sample representation, the key variables and characteristics of interest may vary depending on the context of a particular program. For example, representation of certain groups may be more significant in certain fields (such as women in STEM fields).

Examples

The following examples provide illustrations of different sampling strategies and how to determine sample size.

Student Papers

A department wants to use papers from their writing course to consider their communication SLO. The course has about 100 students over the course of a year, and each student completes a final paper. The department does not have the time and resources to evaluate all 100 papers, so they decide to take a random sample of papers. They decide that they can accept a sampling error of 10% with a 90% confidence interval and need a sample of at least 40 papers. Since they have 7 faculty that have agreed to score papers, the department decides to have each faculty member score 6 papers for a total of 42 papers (to allow for unexpected problems). To help make sure the sample is representative, they use a computer program to randomly select 42 papers from their total number of 100 papers.

Student Posters

As part of a 400 level course, all of a program's 150 graduating seniors complete a research poster presentation. The program would like to use these posters for program assessment and have developed a rubric that evaluates students according to program-level learning outcomes. Since the program offers degrees on two campuses (Campus A has about 67% of seniors and Campus B has 33%), they would like to be able to disaggregate the results by campus and feel confident that the sample is representative. Consequently, they decide on a stratified random sample, and plan to take two-thirds of their sample from Campus A and one-third from Campus B.

The program uses the smallest population size of the two groups (50 seniors at Campus B) to consider the acceptable confidence interval and sampling error rates. They decide to assess 35 seniors at Campus B, and because Campus A has twice as many seniors, they assess 70 seniors from Campus A.

Survey

A department decides to use a senior exit survey to get a sense of student perception of their degree program. Over two semesters, the department has about 400 graduates. They decide to administer the survey electronically to all 400 graduates and receive responses from 196 students (a 49% response rate). While the department is pleased with the sample size for the survey, they are concerned that certain groups of students may not have been motivated to respond and that the results may not be representative of all students.

While the survey was distributed anonymously, it included a set of demographic questions (e.g. option, campus, sex, and first-generation status). The department decided to compare the distribution of responses from these questions to demographic information about all 400 graduates, and found that the respondents typically paralleled the entire group of graduates in terms of option, campus, sex, and first-generation status, helping to alleviate the concern.

Focus Group

A program decided to conduct a focus group with a group of seniors nearing graduation to ask questions about students' experience in the curriculum and their confidence in particular skills. The program invited all 20 seniors in the capstone course to participate in a focus group conducted by one of ATL's assessment specialists (a neutral 3rd party), and 12 seniors showed up to participate. Since focus groups are by nature semi-confidential (i.e. participant names are not collected nor provided to the program), it is often not possible to determine how representative the sample is. Keeping in mind that focus group results are often suggestive, rather than definitive, the program decides to use the results in conjunction with other sources of evidence.

References

- Dillman, D.A. (2007). Mail and Internet Surveys: A Tailored Design Method, 2nd Ed. New York, NY; John Wiley & Sons, Inc.
- The Pell Institute for the Study of Opportunity in Higher Education. Evaluation Guide. http://toolkit.pellinstitute.org/evaluation-guide/
- Suskie, L. (2018). Assessing Student Learning: A Common Sense Guide, 3rd Ed. San Francisco, CA; Jossey Bass.