PPI CLASS Rating and IDM Survey Sampling Plan

Cultivate Learning
College of Education
University of Washington

February 27, 2019

Executive Summary

Cultivate Learning in partnership with state Research Partners in line with the CQI improvement strategy of the Bill and Melinda Gates Foundation will work to populate the Implementation Development Map. This map will host information on high quality teaching (CLASS), professional development, curricula, and child assessments which will be obtained through CLASS ratings and surveys. The intention is to populate IDM indicators using quantitative data collected at the classroom level. To this end representative sample of CLASS ratings will be collected and surveys will be sent to the classrooms in our focus states to ask about the information to populate the IDM. This document provides guidelines for the Research Partners outlining priorities for this process and recommendations for the execution of the data collection protocols. An example power analysis is shown in this document that illustrates how to decide on the sample size – how many classrooms to be rated on CLASS and how many surveys to send out. Not all data necessary for such an analysis is available to us at this time (data sharing agreements in progress) so this analysis will use some approximations and the Research Partners can use their own data to adapt this procedure to their states.

Contents

1 Goals

Within the PPI project we would like to sample state Pre-K classrooms, determine their CLASS ratings and survey staff about elements of the IDM – professional development, curriculum, child assessments, and high-quality teaching. We would like to do this annually over the lifetime of the PPI grant (2021) so that the IDM self-assessment is populated with data that give stakeholders an indication regarding the progress toward the goal of implementing high-quality Pre-K in participating states based on the 15 Essential Elements strategy pioneered by the Gates Foundation.

We would like to design a sampling plan that would be *representative* in terms of:

1. size of the program
2. whether program is located in rural or urban area
3. whether the program is run by the school district or community (tribal, religious etc)
4. ethnicity of the children enrolled in the program
5. whether children are dual language learners

The survey presents an additional challenge in terms of the response rate, which might require appropriate incentives to guarantee a sample of reasonable size and reasonably representative in terms of the dimensions/criteria above. The following outlines the steps in the process of accomplishing these goals and the following chapters in this document provide details on how these steps may be accomplished and how we want to make sure this effort is coordinated across focus states and what are the parameters within which we would hope the Research Partners remained when implementing this plan and what sort of details should be described in their sampling plans. All of these steps are to be completed by June 30th.

1. decide whether a new sample will be collected or whether existing data can be used, if existing data is used, we encourage the Research Partners to make sure this sample is representative and if not to either collect additional data to make it representative (if some strata are completely missing) or to use appropriate weights and supply us with those weights when providing the data, so that when we compute mean and other statistics we could compute a mean corresponding to the population mean (if all strata are present, however, some are undersampled with respect to their incidence in the population)
2. choose a sampling design – simple random sampling, stratified sampling, or cluster sampling
3. put together a sampling frame – list of all classrooms in the state Pre-K program, tally them up since power analysis will require this information
4. look for data that can be used to inform the power analysis (similar data collection performed in the past or by other state or research organization that could provide means and standard deviations of the relevant variables to be used in simulations or which can be plugged into a sample size formula)

5. perform a power analysis or compute the sample size required using existing data on surveys similar to the IDM survey and CLASS ratings

6. decide on a strategy to limit possible non-response or how to take into account after it happens (collect some more data regarding the non-responders and responders so that the two groups may be compared to see whether the non-response is negatively affecting the representativeness of the sample)

7. take the IDM questions in Appendix C and craft good survey questions out of them with the emphasis on such questions still answering the IDM indicators of interest and eliciting good responses from the target audience

8. harmonize the protocol for administering the CLASS ratings so that aspects such as number of cycles or duration of the CLASS observation are consistently applied across all classrooms sampled

9. harmonize the survey administering procedure so that context does not affect responses too much (if questions tacked on other survey already being administered look into whether the preceding questions put respondent into a different frame of mind and significantly affect responses as a result)

10. administer the survey and CLASS ratings

11. create a codebook describing the variables provided in the CLASS rating and IDM survey data so that the data can be properly used by researchers not involved in the data collection process (Appendix B is a good example of a codebook)

12. provide the collected data along with the codebook to Cultivate Learning in an appropriate format (CSV file preferred but is negotiable)

2 Appropriate Sampling Strategies

2.1 Choosing a sample

In this section we will describe what we consider an appropriate sampling protocol. This information can also be accessed in Lohr (2009) and Lumley (2011). Stating this at the outset is important to make sure that the information that we collect really communicates what we think it communicates. Lohr (2009) puts it this way: "Surveys and samples sometimes seem to surround you. Many give valuable information; some, unfortunately, are so poorly conceived and implemented that it would be better for science and society if they were simply not done." Lohr (2009) uses SAS for all expositions and Lumley (2011) uses the R software, which we will use in this document as well. Of course, Research Partners are free to use any software they deem desirable. We will reproduce the main points articulated in Lohr (2009) here briefly and more details are available in Appendix D of this document.

There are several decisions one has to make when setting out to collect a sample:

- observation unit – in our case we have to decide whether we are looking at classrooms, programs, or sites
- target population – in our case we have to define what constitute a state Pre-K program or classroom, this might seem trivial at first but the fact that many classrooms use braided/blended funding makes this less than trivial, one possible definition is "every classroom with at least one child sponsored by the state Pre-K program"
- sampling frame – in our cases this would constitute a list of all observation units, the whole target population; a sample would then be selected by choosing from this list (this information is unavailable to us at Cultivate Learning at the moment, so we leave this to the Research Partners)

2.2 Selection bias

Selection bias is an issue that will be relevant for the IDM survey questionnaire and most likely not for the CLASS rating sampling. To hear it from the expert's mouth: "Selection bias occurs when some population units are sampled at a different rate than intended by the investigator" (Lohr (2009)). In our case, for example, we may decide to use simple random sampling of the classrooms. We put together a list of all classrooms in the state that host at least one child financed through the state Pre-K program (OPK/PP, VPK, ECEAP) and randomly choose a sample. Subsequently, we go ahead and send surveys to that sample (to the teachers of those classrooms for example). Let us assume that we sent out the survey via FedEx and FedEx happens not to delivery to native American tribal areas (some of which contain classrooms on our sampling frame list). Let us further assume that there are 1,000 classrooms in the state and we decide to survey 100 of them. Then every classroom should have in theory a probability of $1/10$ of being sent a survey. However, because of the aforementioned issues with FedEx not delivering to tribal areas, the classrooms in tribal areas would have a probability of 0 of being sent a survey. This different between intended probability of 10% and actual probability of 0 is what is referred to as selection bias. In

this particular case, the sample would be biased towards non-tribal classrooms. As we outlined in Section 1 we do not want this to happen because we want the sample to be representative of those areas. Hence, selection bias is undesirable for us and we want the Research Partners to look out for it.

Convenience sample "A sample of convenience is often biased, since the units that are easiest to select or that are most likely to respond are usually not representative of the hrader-to-select or nonresponding units" (Lohr (2009)). A convenience sample is a sample that was not prepared in any statistical way and where we do not know what is the probability of any particular unit or group of units of being in the sample. It is called convenient because it is collected simply by obtaining whichever units we were able to obtain. Administrative datasets are common examples of a convenience sample. In our project, a good example of administrative dataset that is a convenience sample is the annual sample of ECEAP class ratings. In WA state every ECEAP program has to be rated within a three year period. Which program is rated in any particular year depends on whether it has been rated previously, whether it is ready to be rated, what is the current DCYF budget, etc. This does not necessarily mean that such a sample would be biased, however, in such a case one needs to explicitly investigate whether the sample is biased and we would like to encourage the Research Partners to include a description of such an investigation in their sampling plans.

Judgement sample This one is little confusing but an important distinction to make. When we talk about a representative sample, what we mean is that we use statistics to understand the overall population and statistics to assess the sample, or we use a statistical procedure that is guaranteed to give us a representative sample even if do not know the population's characteristics. Judgement sample takes place if we remove the statistics and try to collect a sample of representative classrooms by common sense, intution or some other heuristic not based on statistics. For example, if someone says that classrooms in Pierce county are "pretty representative" of what a typical classroom looks like in Washington state, proceeds to only survey Pierce county classrooms and then claims that this is a representative sample without any evidence and without using statistics to arrive at the original conclusion that Pierce county is representative of Washington state, then this is called a judgement sample and as human judgements tend to be biased (Tversky & Kahneman (1974)) so would the sample collected in such a way be biased.

Nonresponse As we mentioned above, non-response is a big issue for survey sampling and would be an issue for the IDM survey while not necessarily for CLASS sampling. Why is non-response a big issue? Because it could potentially give rise to a biased sample. If there is any reason why some classrooms are more likely to respond than others, then this would introduce an unintended difference in sampling rates of those classrooms (note that this was our definition of bias). Let us say (this is fictitious example and not based on facts) that tribal communities have to pay more for mail and that we asked them to mail the survey to us after it is filled. Then these communities would be less likely to respond and this would lead to their smaller than intended probabilities of being in a sample and the sample would then be biased towards containing more non-tribal responses/classrooms. This goes against our goals of having a representative sample (Section 1).

To guard against this issue, we want the Research Partners to collect information about the classrooms/sampling units that did not respond to the survey and either demonstrate that they are similar to the ones that responded (based on the representativeness criteria described in Section 1)

or estimate how different the responsive classrooms are from the non-responsive ones and adjust the estimates accordingly after the fact. This would require using state databases on state Pre-K contractors (for example, in Washington state ELMS would be such a database).

2.3 Types of Random Sampling

An effort to collect a representative sample on the data (CLASS and survey data) puts us in a position where we need to use probabilistic sampling to accomplish our goal (rather than using convenience or judgement sampling which would not results in a representative sample). Lohr (2009) defines probabilistic sampling in the following way: "In a probability sample, each unit in the population has a known probability of selection, and a random number table or other randomization mechanism is used to choose the specific units to be included in the sample." In other words, we know at which rate certain groups were sampled because we intended it so and made sure our intentions were properly implemented. The types of sampling that satisfy these criteria are *simple random sampling, stratified sampling, cluster sampling, and systematic sampling*. All of these approaches would guarantee a representative sample and their are all acceptable for the CLASS sampling and IDM survey sampling plan in our focus states. We would prefer to have these approaches be as uniform across focus states as possible, however, it is not strictly necessary. As long as statistically sound approach is followed, a representative sample should arise.

Simple random sampling In a random sampling, every unit has the same probability of appearing in the sample:

$$P(\text{being sampled}) = \frac{n}{N} \tag{1}$$

where n is the size of the sample and N is the size of the population.

The execution of this sampling scheme is quite simple, it is also easy to communicate the procedure to stakeholders and other interested parties. One puts together the sample frame (list of all units in the population, in our cases it could be classrooms) and then randomly select a given number of these units to be observed (CLASS) or surveyed (IDM). It is important that an actual random draw is used (ideally using statistical software) to select the sample from this population rather than leave this as an adhoc decision to administrative staff, as this could result in a convenience sample, which would most likely have less power than the random sample as well as most likely not be representative of the state population of Pre-K programs (see Section 1).

Stratified random sampling This is the approach that Tennessee is currently taking with CLASS sampling. Stratified random sampling has the potential to reduce sampling variation and so could save money by achieving the same power (statistical) with less data. The procedure for stratified random sampling is similar in most respects. One has to obtain a list of all programs in the population (all eligible programs). These programs would then be classified into the individual strata. Within these strata a certain number of programs would be randomly selected. The number of programs selected would be equal to the overall size of the sample times the proportion of the strata programs in the overall population of all Pre-K programs if no oversampling is considered and variances are the same in all strata.

One has to decide what the strata would be in stratified sampling and we would prefer if those categories were some of the criteria outlined in Section 1, however, even if they were not (in Tennessee stratification is based on VPK application scores, a type of classroom quality measure) the sampling is still done by random within the strata and so should result in a representative sampling along the categories listed in Section 1 even if those do not coincide with the actual strata used.

If we believe that the within-strata variance of CLASS ratings differs across strata in a meaningful way, we can oversample the strata where CLASS ratings have higher variance. The number of programs sampled from a given strata is proportional to the overall number of programs in the strata and their variance in the following way:

$$n_k \propto N_k \sigma_k \tag{2}$$

where N_k is the size of the strata population and σ_k^2 is its variance. This is called Neyman allocation after its inventor (Neyman (1934)). These strata samples also have to add up to the overall sample:

$$n = A \sum_{i=1}^{K} N_k \sigma_k \tag{3}$$

where A is determined from our power analysis from our desire to be able to detect a certain fluctuations in year-on-year CLASS ratings. A Monte Carlo procedure will yield the optimal sample size n (see Lumley (2011) for more details).

Cluster sampling Cluster sampling is the most complicated of the three approaches. It does not decrease sampling variation. On the contrary, it increases it. At the same time, it allows possible fuel economies by sampling more units in a geographical area where some units are already sampled. These units will be correlated so more than usual amount of data would need to be collected compared to simple random sampling. This sampling approach might be appropriate if the research partners find that the classrooms within the same contractor have highly correlated results.

Lohr (2009) and Lumley (2011) provide more details on all these sampling approaches. The general rule is that more complicated sampling approaches require more data to execute. For example for stratified sampling it is appropriate to find what are the standard deviations of similar surveys in the different strata before making a decision regarding how many units to sample in each strata (part of the power analysis). When Research Partners decide for any of these approaches, we would ask that they specify the details of those approaches in their sampling plans – what approach was used, how are the strata defined, what are the primary and secondary sampling units, how was the power analysis done etc. Research Partners are encouraged to discuss within the group regarding best practices etc. Cultivate Learning is more than happy to participate in such discussions.

2.4 Working with existing administrative data

Situation may be more complicated if some data already exist that was not collected using any of the sampling schemes mentioned above, but rather using convenience sampling. This is for example the case for the Washington state CLASS sampling effort. In such a case, it is important

to go back to the population (if possible) and try to understand the population composition along the lines outlined in Section 1. One can then compare using a t-test or similar procedure whether the administrative sample resembles the population of the state Pre-K classrooms. If yes, we are done. If not, we need to understand how severe certain misrepresentations are. Are certain strata (DLL children, community-based organizations etc) missing completely or are they merely underrepresented. If they are missing completely, effort needs to be made to collect data from them. If they are merely underrepresented, one needs to create a set of weights to accompany the data, so that representative statistics can be computed from the sample. The notion is that if community based organizations are underrepresented, than they would have higher weights because every community classroom in such a case represents more population classrooms than is the case for non-community based classrooms in the sample. These issues need to be detailed in the Research Partners' sampling plans and documented in the codebooks accompanying the data to be provided to Cultivate Learning by June 30th.

2.5 Survey Questionnaire design

Formulating the questions for a survey is not an easy task. We start with IDM indicator definitions and then translate those into questions. We have done this and Appendix C shows the result. This is a great first step. As the following step, we encourage the Research Partners to look at these questions and see whether they need to be translated or reformulated so that the respondents understand them well and provide good responses. Cultivate Learning Research Coordinators have excellent understanding of the IDM and our Research Partners have excellent understanding of research methodology and we believe it will take both of these to craft good survey questions for the IDM survey. Cultivate Learning will make Research Coordinators available to the Research Partners to clarify intent of the IDM questions. We already received some great comments from the Oregon and Washington Research Partners.

Lohr (2009) has several recommendations regarding survey questionnaire design (see Appendix D for more details; Appendix B provides an example of a survey that follows these recommendations – National Survey of Early Care and Education administered by the Administration for Children and Families):

1. **test your questions before taking the survey:** Questions can be tested on a small sample of members of the target population (say 5). Try different version for the questions and ask respondents in your pretest how they interpret the questions. In our particular case, the IDM contains a lot of policy jargon, we need to make sure that this is translated into a language that our respondents (teachers, instructional leaders) understand. Lohr (2009) also cautions that the questions should really be tested on real respondents, rather than co-workers or other researchers, since co-workers may not have the same understanding of words as persons in our target population.

2. **keep it simple and clear:** Questions that seem clear to the author may not be clear to a person with a different native language, cultural or ethnic background etc. Lohr (2009) cites examples where respondents misunderstood even seemingly obvious words such as "proportion", some people misinterpreted "how long" as "which" etc. More complicated sentences make it more likely that misinterpretation occurs. This reaffirms the need to pre-test the survey questions before distributing them to the target sample.

3. **Use specific questions instead of general ones, if possible:** The policy world is full of vague/abstract phrases (e. g. "instructional leaders", "infrastructure", "capacity") whose meaning is not commonly agreed upon. One has to be careful to avoid these when constructing survey questions. "data-informed processes" is one possible example of a policy phrase that may mean something to a policy-maker or researcher but may be hard to interpret by teachers or instructional leaders and at the end of the day could mean pretty much anything. One can see that for example the NSECE survey in Appendix B gives specific examples or reformulates jargon if jargon's inclusion is necessary. Question 21 in Appendix C is currently phrased "Does your program implement all of the state's comprehensive learning and development standards for Pre-K?." Do teachers and instructional leaders have a good idea when we present them with this question what "all of state's comprehensive learning and development standards for Pre-K" are? If so, we are good to go. If not, we may have to rephrase the question, provide some examples, refer to specific standards etc.

4. **Decide whether to use open or closed questions:** An open question allows respondents to form their own response categories; in a closed question (multiple choice), the respondent chooses from a set of categories displayed. A closed question is a double-edged sword, it will help the respondent with recall but there is a danger it would push them into saying something they would not have said in an answer to an open question. Since our survey is not exploratory, closed questions would make more sense. It is also good to give the respondent a reasonable set of options. Referring to the Question 21 in Appendix C again it is hard to see who will respond "no" to the question, while if options such "most of the time", "always" (more nuanced set of responses) would probably provide more variation in response and more accurate data.

5. **Avoid questions that prompt or motivate the respondent to say what you would like to hear:** These are called leading, or loaded, questions. Questions 21 above is again a good example. If the survey is administered by the DCYF and the question asks whether the respondent is asked in general whether they follow standards, they may feel obligated to say yes, because they are worried to say otherwise to the authority that regulates their organization/business. In a similar vein, Lohr (2009) cautions **"consider the social desirability of responses to questions, and write questions that elicit honest responses"**. This is especially relevant in our case, since as mentioned the survey will be seen as coming from the regulatory authority.

6. **Use forced-choice, rather than agree/disagree questions:** Survey research has shown that people tend to agree with almost anything if it is framed as a agree/disagree question when in fact they have an opinion contrary to statement provided. One of the behavioral biases is that people try to agree with an opinion once this opinion has been articulated to them. So it is better to elicit an opinion from them without providing them with a clue as to what our opinion is. Question 21 is somewhat similar to this situation, but instead of agree/disagree the response options are yes/no. It is worth considering whether a broader set of responses would elicit a more objective opinion of the respondent.

7. **Pay attention to question order effects:** Lohr (2009): "If you ask more than one question on a topic, it is usually (but not always) better to ask the more general question first and follow it by specific questions." For example, if we ask teachers first a general question about professional development opportunities, they will provide their general assessment of how good the opportunities are at their workplace. Then we ask them about specific professional

development program, say from Child Care Aware, and how they benefitted from it, they will recall that and respond. If we switched the order of the questions it is very likely that when they are answering the general question about professional development opportunities the availability bias (Tversky & Kahneman (1974)) will make them think specifically about the CCA training and even though they are asked to answer a general question about professional development, they will still be thinking about that CCA training and their answer will really be another specific assessment of that training masked as a general response. It is also important to be consistent across the focus states in terms of the order of the questions so that we get a consistent set of data from all states.

8. **Ask only one concept per question:** Lohr (2009): "In particular, avoid what are sometimes called double-barreled questions, so named because if one barrel of the shotgun does not get you, the other one will." For example, a question "Do you agree with DCYF's curriculum standards?" is really a two part question, with the first part being whether one agrees with DCYF and the other about curriculum standards. People who have positive opinion of DCYF will tend answer yes regardless of the actual question (somewhat related to the previous point above). Better to drop the reference to DCYF and ask specifically about the curriculum standards.

3 Informational requirements for finalizing the sampling plan

For a proper sampling plan the information required is the following:

1. definition of state Pre-K classroom
2. number of state Pre-K classrooms (a given contractor may operate a large number of classrooms, this needs to be taken into account)
3. definitions of the strata (if stratified sampling approach is employed)
4. definitions of clusters if cluster sampling is used, contractor would make sense as a clutering variable
5. information about the range of change that is reasonable to expect year-on-year in CLASS ratings and in responses to the IDM survey to know the "effect size" to calibrate the power analysis, or what sort of range of changes (or fluctuations) we are expecting over the lifetime of the PPI strategy
6. information about standard deviation in CLASS ratings and correlation between CLASS ratings from two consecutive years, mean and standard deviation of year-on-year differences in CLASS ratings (to calibrate the power analysis), information about standard deviation of responses to the survey questions (this may be challenging to obtain if survey has not been administered before in a given state so Research Partners in the focus states are encouraged to cooperate and share any pieces of useful data; estimates from the literature are also admissible if available)
7. standard deviations of CLASS ratings in the strata (if stratified sampling employed; do some strata have higher variation in the ratings? Then they should be oversampled), standard deviations of survey responses if stratified sampling is employed for the survey

3.1 Definition of State Pre-K program

For the purposes of this data collection effort and to be consistent across the focus states, we will define a state Pre-K classroom as one that hosts at least one child funded through the state Pre-K program. We would prefer the Research Partners to use this definition, however, an exception may be considered if the situation is documented in the sampling plan and the different approach is well justified.

3.2 Definitions of the strata

We know that we want to stratify by a) size of the program, b) rural/urban continuum, c) school district vs community, d) ethnicity, e) DLL (language). The rural/urban continuum could be determined from a zip code so this is not an issue, however, how do we operationalize the other categories?

a) Definition of Size

- Do we use number of children as a basis for the definition of size?
- Do we look at the contractor the classroom belongs under and say that a classroom is "large" if it belongs to a "large" (having many classrooms) contractor?
- Once the above are answered, what is the threshold for saying that a program is "big", eg if children used as basis how many children would constitute a "big" classroom? How many classrooms would constitute a "big" contractor?
- How many sizes do we want? Eg two – big/small? Or three – big, small, medium? Etc.

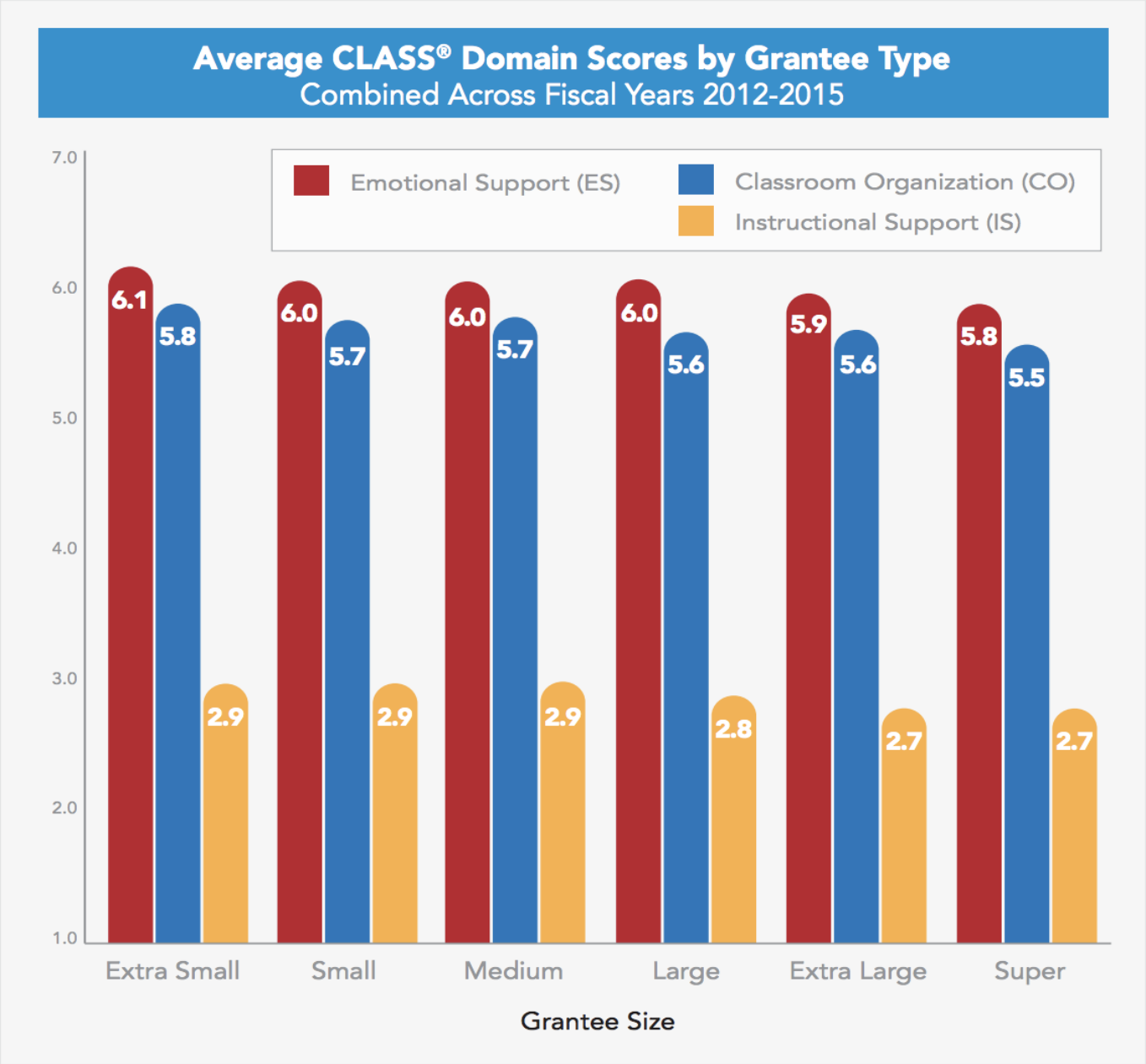
Report on Head Start CLASS Data for Fiscal Years 2012-2015 published by the Administration of Children and Families (ACF) shows there is not much of a difference in terms of CLASS ratings between grantees of different sizes (see Figure 1). So we may have some flexibility in terms of answering the above questions. However, it would be useful for BMGF goals that the answers to these questions are answered consistently across the focus states. We will leave to the Research Partners to attempt to provide answers for the questions above based on their superior knowledge of local circumstances and better access to local data. We would like the decision process to be detailed in the sampling plan so we may understand how these questions were answered and what data was used to answer these questions.

b) Rural-urban continuum The Census has data that assigns every county in the United States to a number that indicates the percentage of the county that is rural. That is one option. This could be a bit crude, since counties are large and ideally one would get a zip code and assignment into whether a given zip code is rural or urban and zip code is small enough area that it should fall entirely under either rural urban, so in this sense it would be more precise. Department of Agriculture also provides information at the county level, so this would not represent an improvement. The Depatment of Agriculture also has that assigning urban-rural continuum to the PUMA – Public Use Microdata Area, which is a geographical designation created by the Bureau of the Census. Each PUMA has a population of at least 100,000 (so that the Bureau can protect the identity of the survey respondents). If county has less than 100,000 inhabitants, then it would be included in the PUMA along with a part of another county to reach the 100,000 threshold. So for example Washington state has 27 counties with populations under 100,000, so county is probably a better unit than a PUMA. Population Studies Center at the University of Michigan has published Urban-Rural continuum codes for zip codes (`http://www.psc.isr.umich.edu/dis/data/kb/downloads/t1101_ziprural.xls`). This would at the moment appear to be the best solution as it would provide consistent definitions across all our focus states. Should Research Partners come across some other way how to identify rural classrooms, we welcome this effort and ask that this approach is documented in the sampling plan. We would prefer for our Research Partners to coordinate determining these aspects of the sampling plan across the focus states, so that the joint PPI effort uses consistent definitions.

c) Community

- We ask the Research Partners to brainstorm the definition of a "community" program. Loosely, what we have in mind is that some classrooms are offered outside of other systems (school

Figure 1: ACF's Study of CLASS Scores and Grantee Size



districts), for example tribal communities, religious communities, or any other communities with some type of autonomy or self-governance/self-regulation. We ask the Research Partners to identify possible communities that fit this notion in their states and try to label each classroom in terms of whether it belongs under a school district or some community of this type. Once such potential communities are identified within the states, the Research Partners and Cultivate Learning can have a conversation regarding how to narrow down or fix the definition of a "community classroom" so that it is consistent across focus states.

- We need exhaustive categories so if there are programs that are not community nor school district-based then we would have to create a "other" category to go with these.

d) Categorizing programs by Ethnicity/Race

- We are interested in categorizing children enrolled in the state Pre-K programs in the following categories: White, African American, Native/Indigenous, Latino, Asian, Pacific Islander/Hawaiian, Other.

- the challenge with this category is that it is really a child level rather than a classroom level characteristic, if state is using administrative data rather than collecting entirely new sample, we would like the Research Partners to collect information on the ethnicity of enrolled children at the child level (only children funded by the state Pre-K program, not all children enrolled in a given classroom) in the population and then comparing the proportions (using a t-test or similar device) of various ethnicities to the proportions of these same ethnicities in the administrative sample. On the other hand, if brand new sample is being collected, all the sampling strategies described in this document should result in a representative sample along this dimension, so this dimension would not enter the data collection plan at the design stage but rather after the data is collected, we would ask the Research Partners to compare the ethnicity proportions in the sample with the ones in the population if possible (if neither Research Partners nor the state has access to the relevant population data, we will rely on the fact that the sampling design was such that it guarantees representativeness). We ask that the Research Partners provide details on this issue in their sampling plans.

- Another possible approach is to collect data at the child level about ethnicity/race and compare children from the sampled classrooms with the children in the non-sampled ones to see if there are any differences in ethnicities represented. If there are, one would reject such a sample and start over (re-randomize). In other words, at the design stage, before one sets out to collect the data, when a sample is drawn from the sampling frame, one performs this check before giving a go-ahead for the actual sample data collection. Alternative approaches can be recommended by the Research Partners and we would ask that they are described in the sampling plan, so that Cultivate Learning may review before Research Partners set out to collect the data.

e) DLL Very similar questions to the ethnicity/race:

- we define Dual Language Learners as children whose first language is not English

- like ethnicity/race this is really a child level characteristic, so it would be complicated to stratify classrooms based on race. Consequently, following our discussion of ethnicity, the suggestions are the same. If state is using administrative data rather than collecting entirely new sample, we would like the Research Partners to collect information on the DLL among enrolled children at the child level (only children funded by the state Pre-K program, not all children enrolled in a given classroom) in the population and then comparing the proportions (using a t-test or similar device) of DLLs vis-a-vis the other children to their proportions in the administrative sample. On the other hand, if brand new sample is being collected, all the sampling strategies described in this document should result in a representative sample along this dimension, so this dimension would not enter the data collection plan at the design stage but rather after the data is collected, we would ask the Research Partners to compare the proportion of DLLs in the sample with their proportion in the population if possible (if neither Research Partners nor the state has access to the relevant population data, we will

rely on the fact that the sampling design was such that it guarantees representativeness). We ask that the Research Partners provide details on this issue in their sampling plans.

- Another possible approach is to collect data at the child level about DLLs and compare children from the sampled classrooms with the children in the non-sampled ones to see if there are any differences in ethnicities represented. If there are, one would reject such a sample and start over (re-randomize). In other words, at the design stage, before one sets out to collect the data, when a sample is drawn from the sampling frame, one performs this check before giving a go-ahead for the actual sample data collection. Alternative approaches can be recommended by the Research Partners and we would ask that they are described in the sampling plan, so that Cultivate Learning may review before Research Partners set out to collect the data.

4 Example power analysis

We would like the Research Partners to include in their sampling plan a detailed description of how they make a decision regarding the size of the sample to collect for the CLASS data collection and survey administration. If Research Partners choose to do universe sampling ("sample" the whole population), no explanation is necessary. The following example for power analysis is specifically tailored to the CLASS data collection plan, however, can be easily modified to apply to the determination of many observations need to be collected for the IDM survey questionnaire as well. We do not have the relevant data for this type of exercise (means, standard deviations of responses to similar survey questions, so we cannot replicate the CLASS power analysis example to the survey, however, it is a matter of changing the relevant variables).

We suggest the following Monte Carlo power analysis procedure (Robert et al. (2010)) may be good way to make a decision regarding sample size if the relevant data to perform the procedure is available to the Research Partners.

As mentioned in Section 3 certain pieces of information are required to perform this analysis. We presently do not have such information available to us, so in the following example we will approximate such information with similar enough information from other contexts.

To operationalize this power analysis, we will use information from data available at Cultivate Learning from other projects involving CLASS data. Research Partners are encouraged to perform this analysis using data available to them. While formulas for optimal sample size exist for simple problems, with complex survey sampling designs it would make sense to use Monte Carlo simulation, because it allows us to calibrate the power analysis closely to the problem at hand. In this approach, we obtain information about the variables in question (see Section 3) – means, standard deviations, correlations – and then randomly draw/simulate the data from the relevant distribution (normal distribution will be used here). This power analysis example is for simple random sampling, since for stratified random sampling, even more information is required (for which we presently do not have even approximations).

For this example, we use CLASS data from the FIND study performed at Cultivate Learning, where repeated CLASS observations were made on a sample of classrooms. We will use this data to determine CLASS means, standard devitions, and correlations between consecutive observations on the same program for the Monte Carlo simulation. Since, this was the best data available at the time, this determined that this exercise will provide power analysis for repeated samplings of the same program annually, which is not a requirement (repeated cross-section is acceptable if deemed more pratical by the RPs). Cultivate Learning will be happy to re-run this power analysis for repeated cross-section if deemed a better alternative by the RPs if the data is available for such an analysis (data on CLASS from two consecutive years from the same state). The estimates of means of standard deviations were corroborated with estimates from the literature (Early et al. (2017) for estimates in Georgia).

4.1 Tennessee

From the NIEER State of Preschool Yearbook and Administration for Children and Families Office of Head Start data we know that enrollment in Head Start is approximately similar in all of our

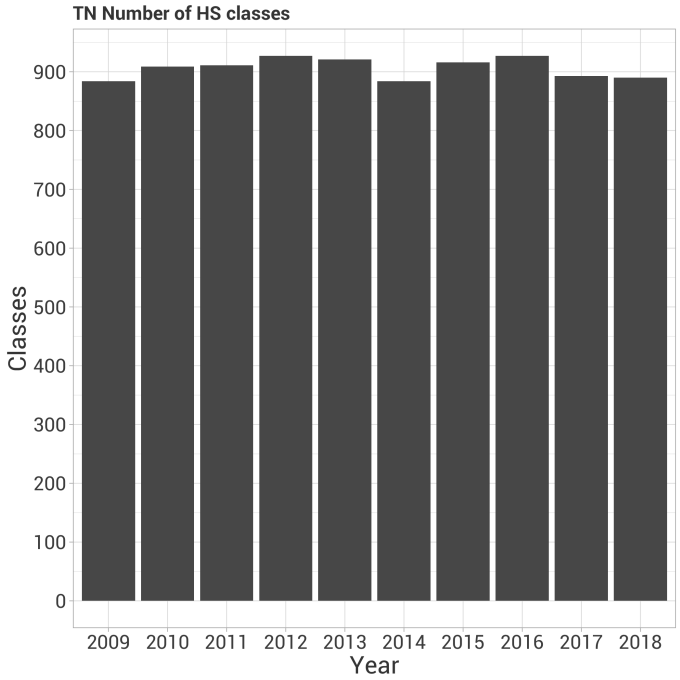


Figure 2: The total number of Head Start classrooms in Tennessee

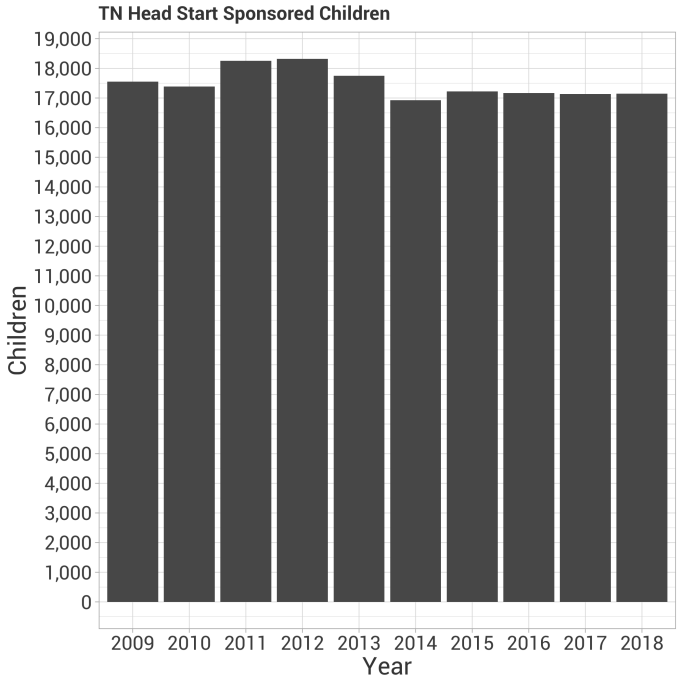


Figure 3: Children enrolled in Head Start in Tennessee

focus states with enrollment in state Pre-K (see Figure 3 and Figure 4). We do not know how many classrooms there are in any of the focus states, but we will use this fact as an excuse to pretend for the purpose of this example to pretend that HS and state Pre-K have the same number of classrooms as well. Tennessee Head Start has around 900 classrooms (see Figure 2).

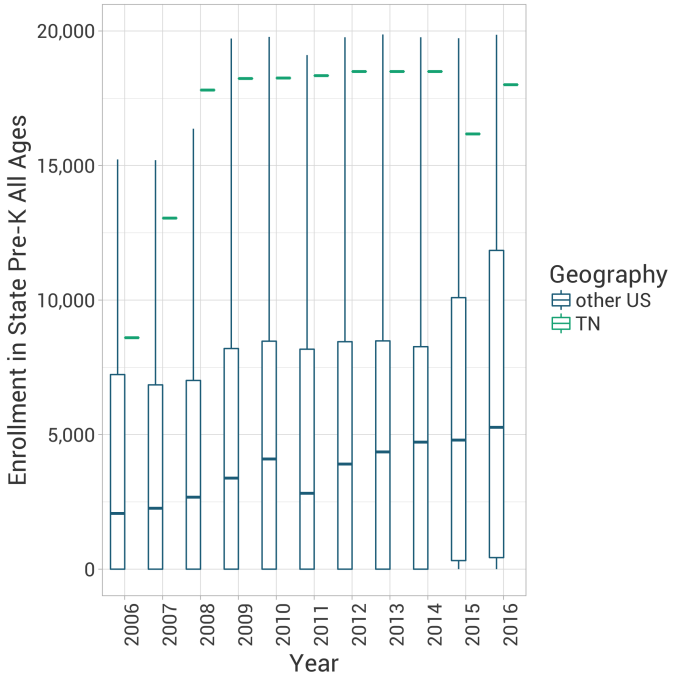


Figure 4: Children enrolled in state Pre-K in Tennessee

We will then assume that thee are around 900 VPK classrooms and so we will incorporate this into the estimation of the sample size required on detect given year-on-year change in the CLASS scores. This information will enter the so-called finite population correction, that takes into account that the non-sampled population gets smaller as we sample from it (Lumley (2011)). As Figure 5 indicates a sample bigger than 125 observations should allow us to detect possible fluctuations in the range of 1-5% about 75% of the time (or confirm the absence of such fluctuations with such level of statistical confidence). A sample bigger than 300 would allow us to detect such changes 90% of the time. The range of 1-5% is functionally similar to effect size. Some choose to perform this calculation by defining what the expected effect size is and then determining how big of a sample would have to be collected to reveal an effect of such size. This is all acceptable and a matter of preference. For the IDM survey sampling we would like the Research Partners to survey enough classrooms so that a change in the range of 10% from one year to the next is detectable 75% of the time. Our Research Partners can replicate this simulation with their own data or Cultivate Learning is happy to perform such simulation if provided with some information such as means and standard deviations (and correlations) for the relevant variables for which the simulation would be performed.

Figure 5 was obtained via Monte Carlo simulation by generating random samples with given properties where the true model had a difference (we do not necessarily expect that the CLASS scores would change every year, however, when they do change at some point in a certain year, we would like to be able to detect such a change) in CLASS scores in the range 1-5% and then we did a hypothesis test on the null hypothesis of no change and the Figure shows how often this null hypothesis was rejected and how often it was not. The proportion of rejection of this (false) hypothesis is the power (red portion of the bar in the figure). The procedure could be broken down into the following steps:

1. decide number of simulations, the more simulations the better the estimates, but 500 should generate reasonably precise estimates
2. simulate samples of different size, decide which sizes we will simulate, in this example we simulated samples of size 50, 100, 125, 150, 200, 250, 275, and 300
3. store information about total programs, means, effect sizes, standard deviations, and correlations
4. draw random samples from the distribution of CLASS scores (assume normal distribution and for mean and standard deviations use the information we stored above)
5. look at the distribution of the random draws, find appropriate quantiles for hypothesis testing (we look at the 97.5th percentile), in this example we look at whether the change is larger than zero in absolute value so we are looking at whether the mean minus the percentile times standard deviation of the class scores are larger than zero. If yes, then we would reject the hypothesis that the change in CLASS scores from one year to the next is zero
6. since we designed the changes to be larger than zero (we choose a number of possible effect sizes in previous step), then we want the null hypothesis to be reject and the proportion of rejections constitutes the power.
7. Finally, we plot the proportion of rejections in Figure 5 to see what the power is.

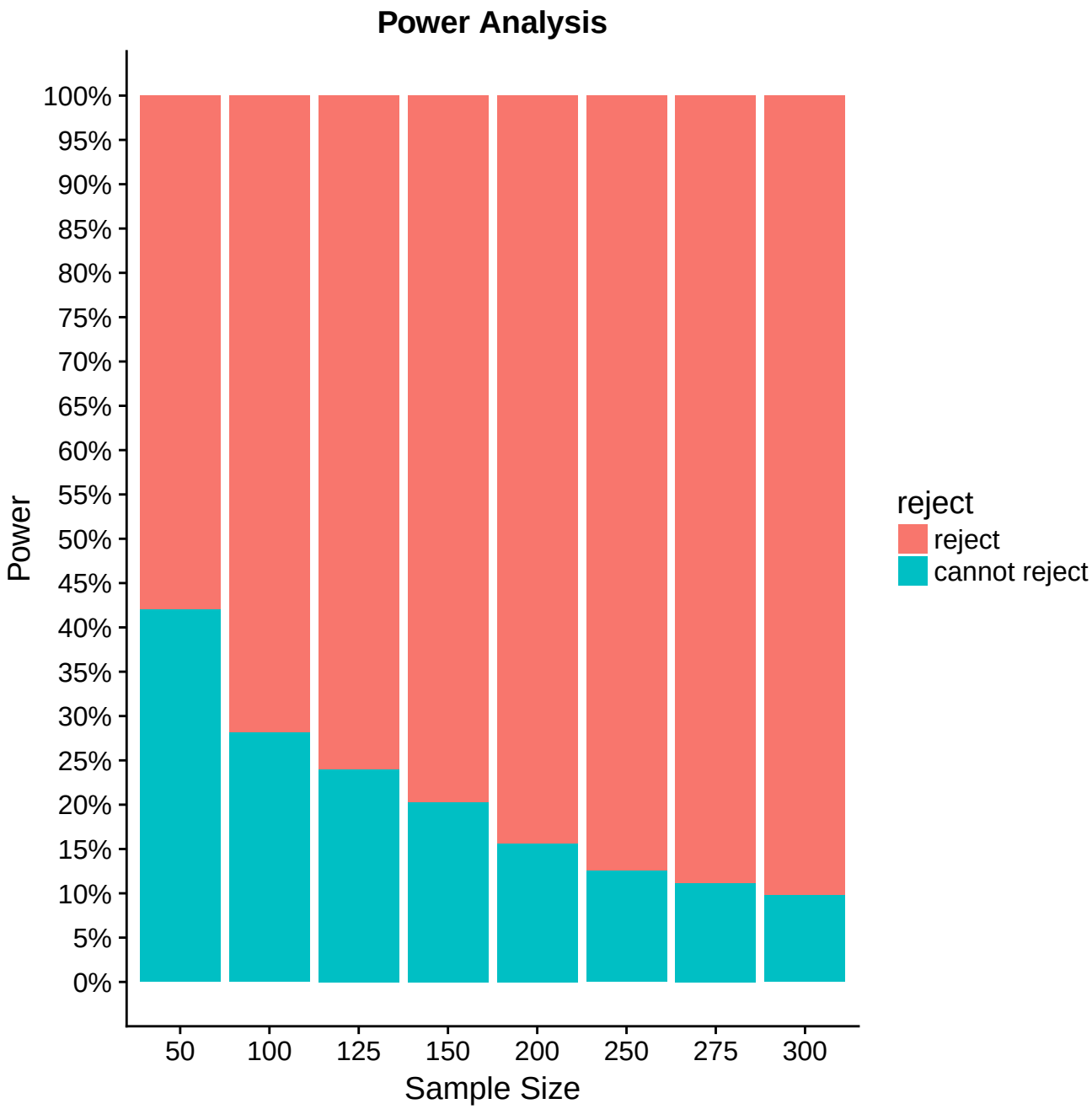


Figure 5: Power to reject the null hypothesis of no difference for various sample sizes (TN)

4.2 Oregon

Figure 6 indicates that there between 800 and 900 classrooms in Oregon's federal Head Start program. Figure 7 and Figure 8 indicate that enrollment is comparable across federal Head Start and state Pre-K (including state Head Start), so we will make the assumption here that there are 850 classrooms in Oregon state Pre-K program (note these are classrooms and not "programs" or "contractors"; after discussing this issue with the Washington Research Partners we have concluded that referring to the contractors as "programs" is somewhat confusing and ill-defined). While it is somewhat surprising that Tennessee and Oregon would have similar number of classrooms, this could be explained by having smaller classrooms or having fewer state Pre-K sponsored children in proportion to children funded from other sources. In any case, the following is just a demonstration/exercise and the Research Partners are encouraged to replicate this exercise with their superior data.

In this example, we change the configuration of the Monte Carlo exercise and use effect sizes of 5-10% and look for a sample size that will give us a reasonable probability of finding year-on-year changes of such magnitude in the data, this is the sort of effect size range that we would like the Research Partners to calibrate their sample size for the IDM survey towards finding. If Research Partners can convincingly demonstrate in their sampling plans that the kind of survey data that we are looking for very rarely shows changes as small as 5-10% but rather usually tends to change more, this may justify their using a different range for sample size estimation. Such argument will require detailing what sort of data was used for this determination and tabulating the data to support this argument.

As Figure 9 indicates, a sample of at least 50 classrooms will allow us to detect 5-10% fluctuations 90% of the time. This will serve us well even if there are no fluctuations since we will be able to state the absence of fluctuations with the same level of statistical confidence. 75% power and the range of detection of 5-10% is acceptable for the purposes of the IDM survey questionnaire sample size. Note that in the figure, the red portion of the bar represents power, so power can be computed by taking the number on the vertical axis at a point where the bar splits between the two colors and subtracting that number from 100% (counting from the top). Note again, these estimates are not to be taken literally, but rather an illustration of the procedure. For simple random sampling, a closed form formula for the sample size is also acceptable. Monte Carlo procedure shown here shines primarily in more complicated sampling designs.

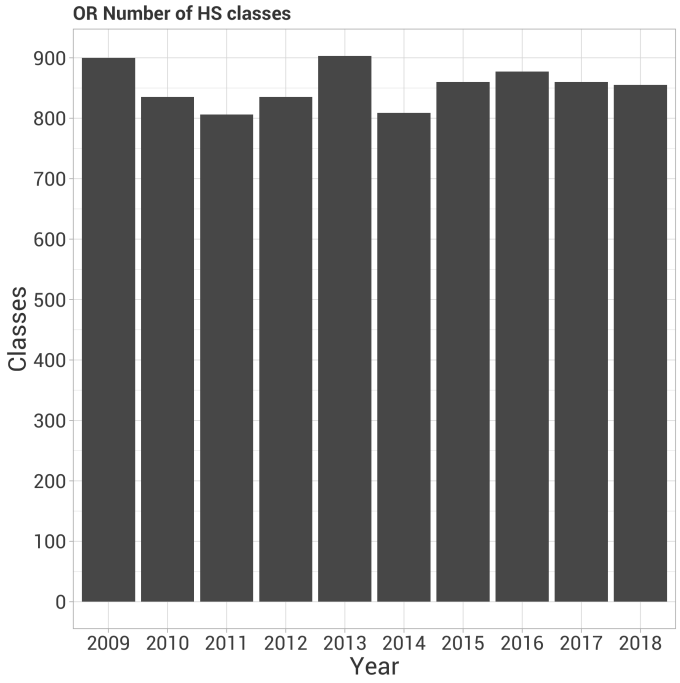


Figure 6: The total number of Head Start classrooms in Oregon

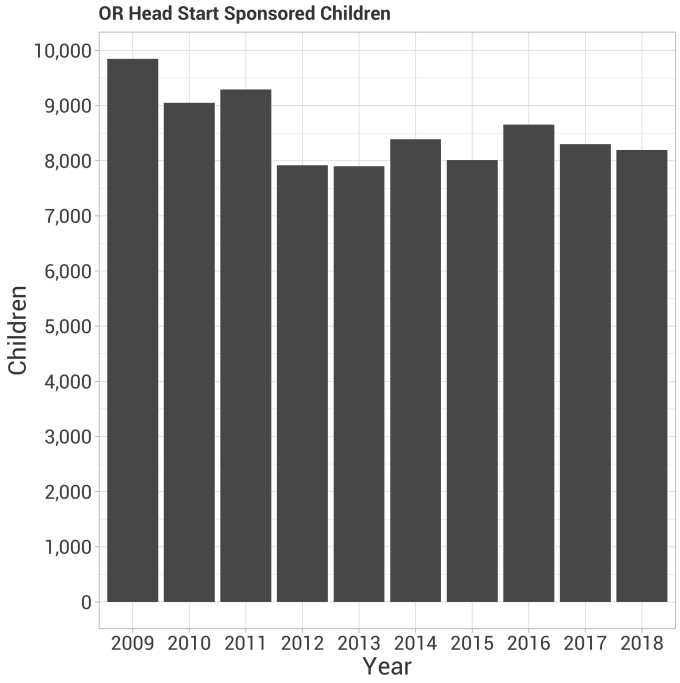


Figure 7: Children enrolled in Head Start in Oregon

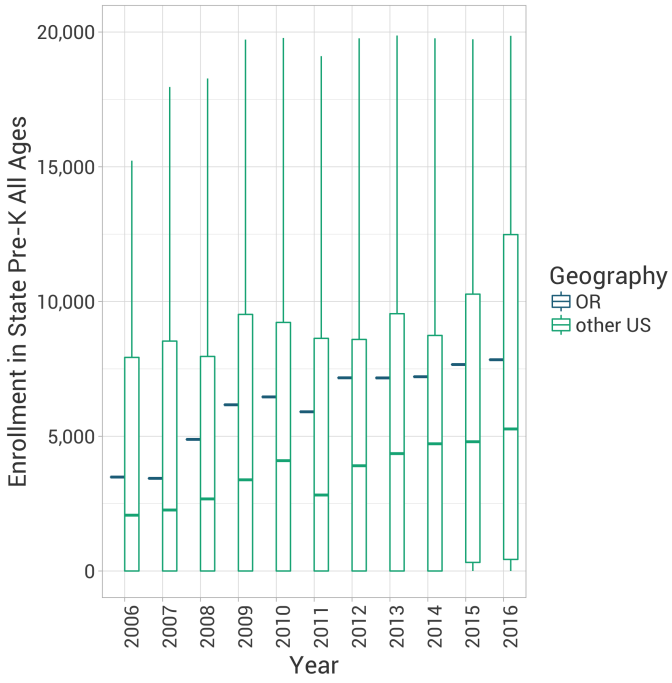


Figure 8: Children enrolled in state Pre-K in Oregon

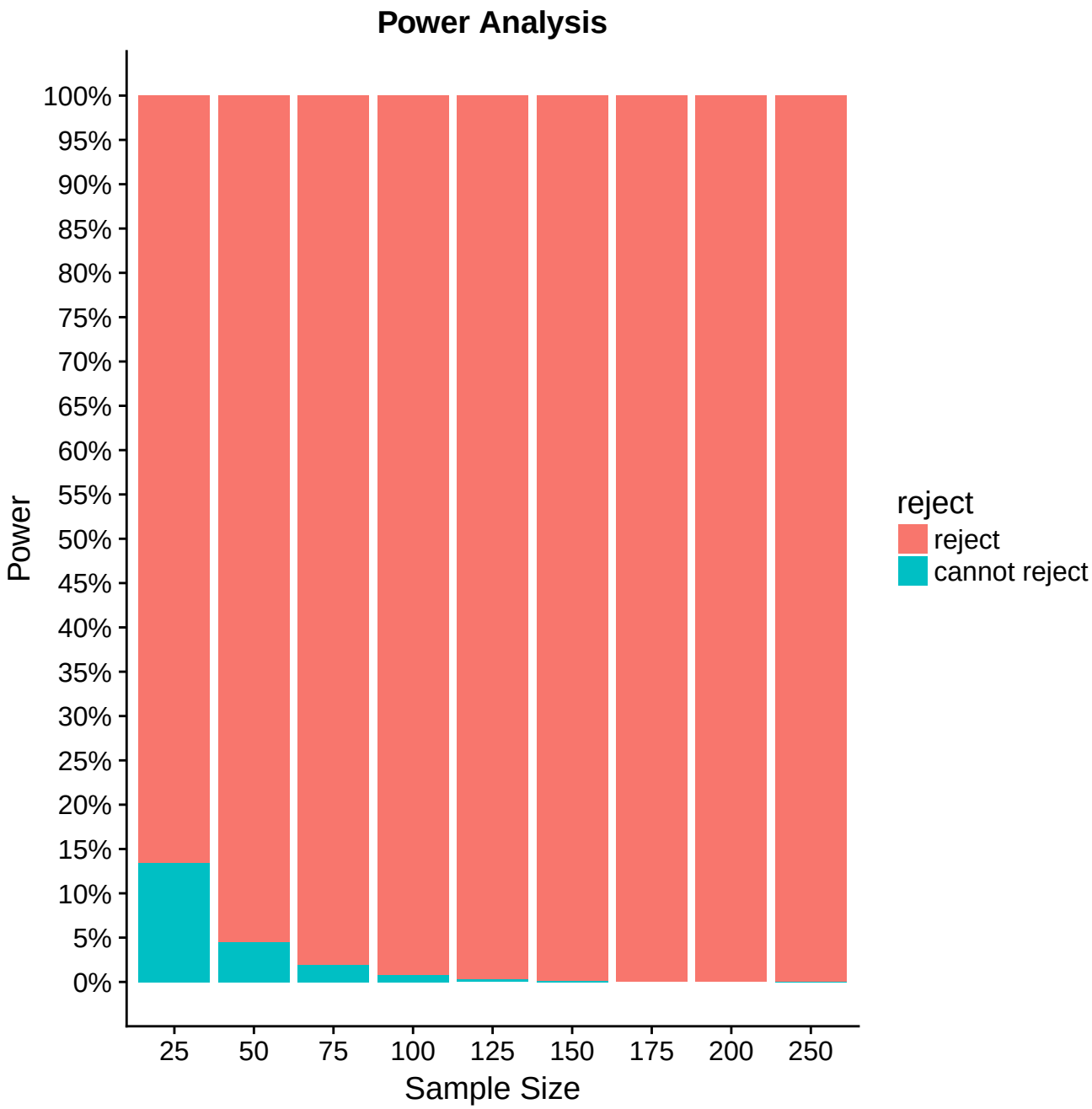


Figure 9: Power to reject the null hypothesis of no change for various sample sizes (OR)

4.3 Washington

Using similar reasoning as we did above we also reach the conclusion that Washington has 850 state Pre-K classrooms (see Figure 10, Figure 11, and Figure 12). This means that our simulation would proceed along similar lines as the one for Oregon did and so for the purpose of demonstration of how the procedure depends on the input parameters, in this example we will change the assumed variation (standard deviation) in CLASS scores by 25%. Once can see in Figure **??** that it now requires 75 observations to reach 90% power to detect the effect sizes in the range of 5-10%. This illustrates the importance of getting good input data for running the power analysis.

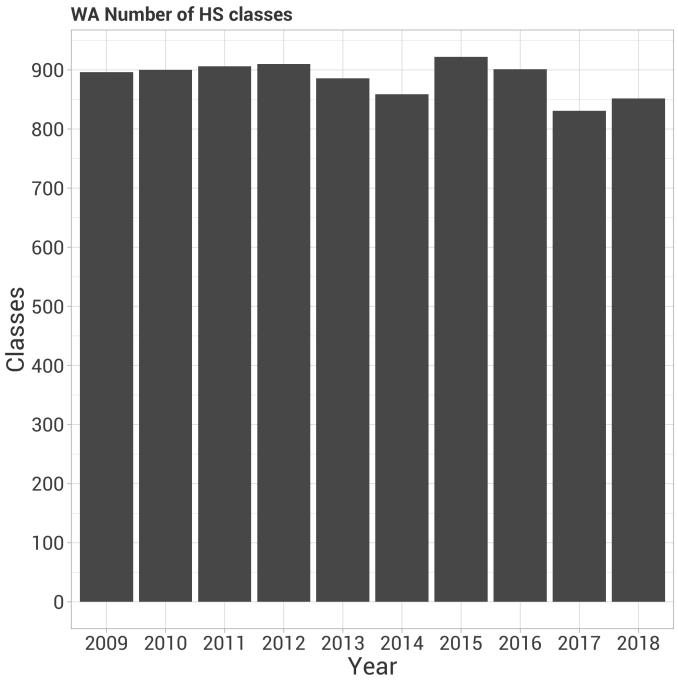


Figure 10: The total number of Head Start classrooms in Washington

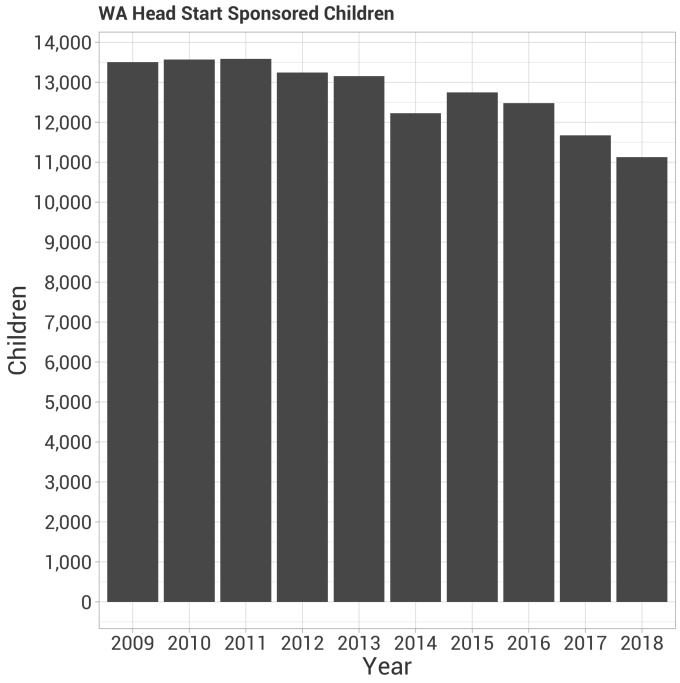


Figure 11: Children enrolled in Head Start in Washington

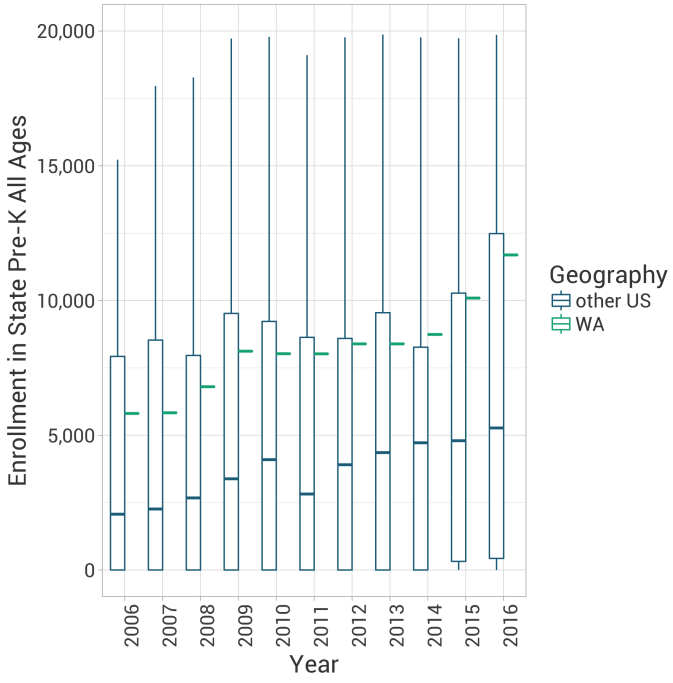


Figure 12: Children enrolled in state Pre-K in Washington

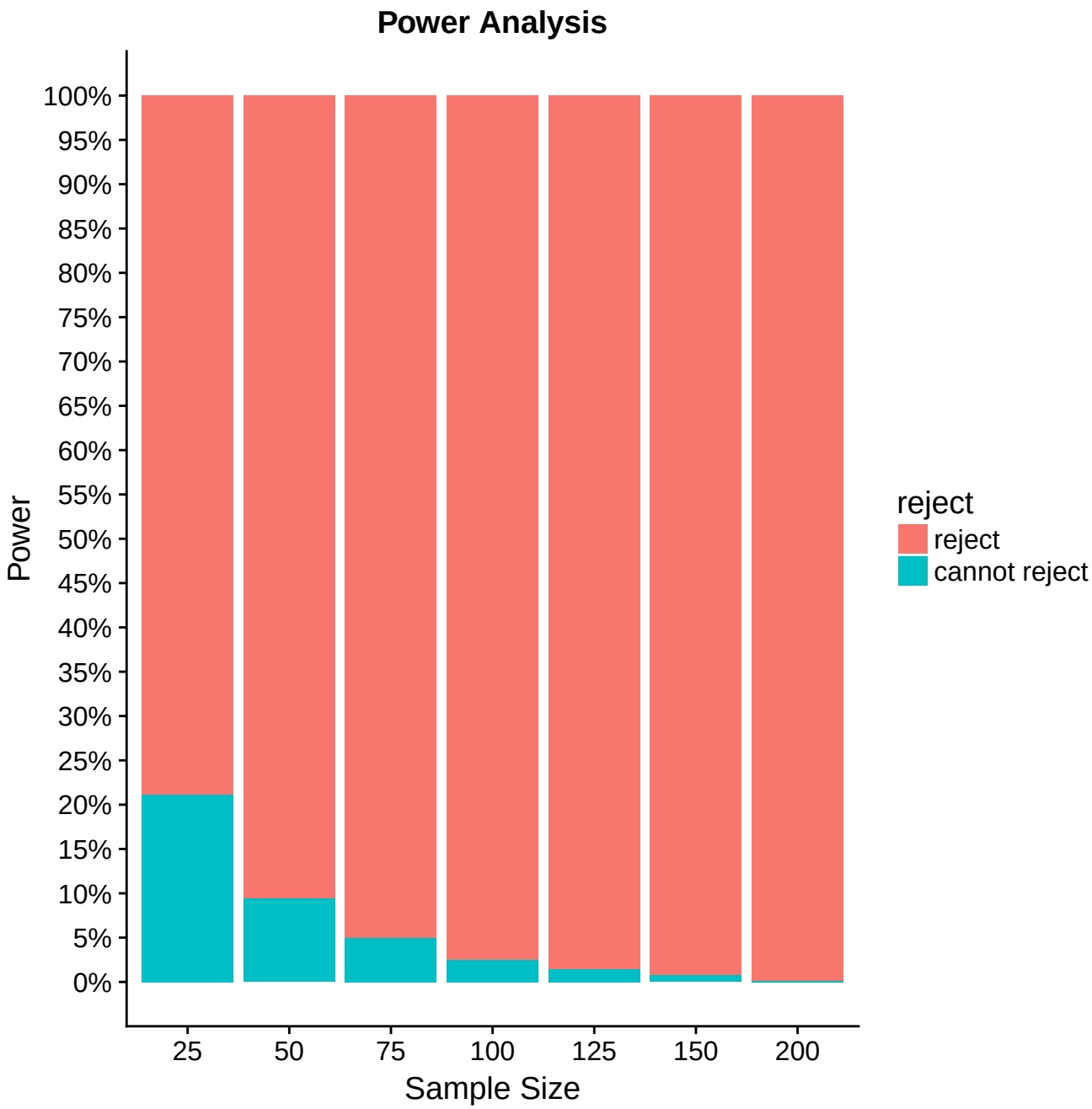


Figure 13: Power to reject the null hypothesis of no difference for various sample sizes (WA)

5 CLASS Rating protocol

Our main observation unit is a classroom, yet classroom are clustered within programs/contractors and to some extent may be similar to each other (outcomes of interest may be correlated highly within this cluster). One has to make a decision regarding this issue. In addition to this, one has to make decision regarding the rating tool itself – how long the observation will be or how many cycles of CLASS are performed. Appendix A is a CLASS rating protocol employed by Cultivate Learning for CLASS ratings contracted for by the DCYF of Washington state. We would like to ask our research partners to detail their CLASS observation protocol and would prefer if such protocol was consistent across observations within the same state and ideally across our focus states. This should not be much of an issue in Washington and Tennessee, however, given that Oregon operates two autonomous state Pre-K programs, we wonder whether this would be an issue and would like to understand how similar the CLASS observation protocols across the two programs will be (PP and OPK). We also encourage the Research Partners to share best practices regarding CLASS observation protocol and if possible employ similar procedures.

6 References

Early, D. M., Maxwell, K. L., Ponder, B. D., & Pan, Y. (2017). Improving teacher-child interactions: A randomized controlled trial of making the most of classroom interactions and my teaching partner professional development models. *Early Childhood Research Quarterly*, *38*, 57–70.

Lohr, S. L. (2009). *Sampling: design and analysis*. Nelson Education.

Lumley, T. (2011). *Complex surveys: A guide to analysis using r* (Vol. 565). John Wiley & Sons.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625.

Robert, C. P., Casella, G., & Casella, G. (2010). *Introducing monte carlo methods with r* (Vol. 18). Springer.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Appendices

A Early Achievers CLASS Rating Protocol

LEARNING ENVIRONMENTS SELECTED FOR OBSERVATION

How are Learning Environments from my site chosen for ERS assessment? Learning Environments are randomly sampled by age group**.

Age groups as defined by the measures are: ITERS-R: 0-30 months, ECERS-R: 30 months-5 years, FCCERS-R: 0-5 years, Infant CLASS 0-15 months, Toddler CLASS 15-36 months, Pre-K CLASS: 36 months-5 years, combined CLASS 0-5 years. When conducting an assessment, the measure is selected based on the ages of the children enrolled in a classroom. When classrooms have children that span the age ranges of multiple measures the measure is selected based on the age group that has the most children enrolled in that classroom (e.g., for a classroom with 4 children in the ITERS age range and 5 children in the ECERS age range, ECERS will be conducted).

Learning environments are placed into one of 3 age groups for sampling, this is independent of the assessment requirements: infant: 0-12 months, toddler: 12 months-30 months, and preschool 30 months-5 years.

How many of my site's Learning Environments will be sampled? 1/3 of each age group will be selected to receive full data collection: ERS, LENA, teacher interview, and CLASS*. See below for specifics.

SITES WITH ONE LEARNING ENVIRONMENT *(per age group)*

Sites with one room of an age group (i.e. Infant, Toddler, Preschool) will receive ERS, LENA, teacher interview, and 4 cycles of CLASS*.

SITES WITH TWO LEARNING ENVIRONMENTS *(per age group)*

Sites with two rooms of an age group (i.e. Infant, Toddler, Preschool) will have one sampled classroom for that age group. The sampled room will receive ERS, LENA, teacher interview, and 2 cycles of CLASS*.

The other room will receive 2 cycles of CLASS* only.

SITES WITH THREE to TEN LEARNING ENVIRONMENTS *(per age group)*

Sites with between three and ten rooms of an age group will have 1/3 of the rooms sampled. The sampled rooms will receive ERS, LENA, and teacher interview. These room(s) will not receive a CLASS* observation.

2/3 of the total number of Learning Environments of that age group will receive 2 cycles of CLASS* only.

SITES WITH MORE THAN TEN LEARNING ENVIRONMENTS *(per age group)*

Sites with more than ten rooms of an age group will have 1/3 of the rooms sampled. The sampled rooms will receive ERS, LENA, and teacher interview. These room(s) will not receive a CLASS* observation.

1/2 of the total number of Learning Environments of that age group will receive 2 cycles of CLASS* only.

*Currently CLASS observations do not occur in Infant rooms.

**In Family Childcare programs where children are in separate groups with different providers and children that rarely come together, a separate observation is required for each group and the protocols above will apply.

5/6/2016

B NSECE Workforce questionnaire

WF_A6_STATECERT (QT = WF_CAREER_CERT)

A6. Do you have a Child Development Associate (CDA) certificate or state certification to teach young children, special education or elementary school?

1. Neither State Certification nor CDA
2. State Certification Only
3. Child Development Associate (CDA) Certificate Only
4. Both CDA and State Certification
5. Don't know/Refused/No answer

A7. In the past 12 months, have you done any of the following to improve your skills or gain new skills in working with children? **(***GRID: question format shown in Appendix 6.1.*)

1. Yes
2. No
3. Don't know/Refused/No answer

WF_A7_WORKSHOP (QT = WF_PROFDEV_WRKSHP)

A. Participated in any workshops, for example, those offered by professional associations, resource and referral networks, etc.?

WF_A7_COACH (QT = WF_PROFDEV_COACH)

B. Participated in coaching, mentoring or ongoing consultation with a specialist?

WF_A7_VISIT (QT= WF_PROFDEV_VISIT)

C. Made visits to classrooms in other programs?

WF_A7_MEET (QT= WF_PROFDEV_MEETING)

D. Attended a meeting of a professional organization (such as Zero-to-Three, Association for Education of Young Children; Association for Family Child Care, National After School Association, or another group)?

WF_A7_COMMCOLL (QT = WF_PROFDEV_COURSE)

E. Enrolled in a course at a community college or four-year college or university relevant to your work with children under age 13?

WF_A7_WORKSHOP_SESSIONS (QT= WF_PROFDEV_WRKSHP_TYPE)

A7a_1. Was that a single workshop or a series of several sessions?

1. Single workshop
2. Workshop series
3. Don't know/refused/no answer

SKIP LOGIC BOX 1

If R answers YES to any of **WF_A7_WORKSHOP, WF_A7_COACH, WF_A7_VISIT, WF_A7_MEET** or **WF_A7_COMMCOLL** then ask **WF_A8_ACTIVITIES.**

If R answers NO or Don't Know/ Refused to all of these, then skip to **WF_A9_PROFASSC.**

WF_A8_ACTIVITIES (QT = WF_PROFDEV_GROUP)

Variables affecting eligibility for this item: **SKIP LOGIC BOX 1**

A8a. Did you participate in any of these activities as part of a group from your program?

1. Yes
2. No
3. DON'T KNOW/REFUSED/NO ANSWER

A8B. During the past 12 months, did you receive any of the following types of assistance with the costs of improving your skills, either from your employer or from a local or state agency, college or university? (GRID: *question format shown in Appendix 6.1.*)

1. Yes
2. No
3. Don't know/refused/no answer

WF_A8B_TUITION (QT = WF_PROFDEV_HELP_TUITION)

Variables affecting eligibility for this item: **SKIP LOGIC BOX 1**

1. Assistance with direct costs such as tuition or registration fees

---

WF_A8B_OTHERCOST (QT = WF_PROFDEV_HELP_COST)

Variables affecting eligibility for this item: **SKIP LOGIC BOX 1**

1. Help with other costs of participation such as travel or child care for your own children

---

WF_A8B_RELEASE (QT = WF_PROFDEV_HELP_TIME)

Variables affecting eligibility for this item: **SKIP LOGIC BOX 1**

1. Release time to participate in the activity

WF_PROFDEV_TOPIC
(R) WF_A8C_MAINTOPIC_R

Variables affecting eligibility for this item: **SKIP LOGIC BOX 1**

A8C. What would you say was the main topic of the most recent activity you participated in to improve or gain skills in working with children? For example, was it focused on health and safety, working with families, preparing children to do well in school, techniques for discipline and classroom management, or some other topic?

> ***CAPI:*** INTW read codeframe, only when necessary.
>
> ***WEB/PAPI SAQ:*** R entered responses, without codeframe options listed. Verbatim responses were later coded back into the original codeframe.

1. Health and safety in the classroom (n=959)
2. Cognitive development, including early reading or math (n=495)
3. Doing well in school, including homework assistance, instruction or co-curricular activities (n=377)
4. Helping children's social or emotional growth, including how to behave well (n=971)
5. Physical development and health (n=173)
6. How to work with families (n=300)
7. Serving children with special physical, emotional or behavioral needs (n=356)
8. Working with children who speak more than one language (n=54)
9. Planning activities that meet the needs of the whole class (n=194)
10. Other (n=36)
11. Added: Multi-topic geared to certification, accreditation, standards/QRIS (n=88)
12. Added: Multi-topic geared to general skills (includes developmentally appropriate practice) (n=123)
13. Added: Degree preparation (n=14)
14. Added: Child protection: abuse prevention, reporting (n=55)
15. Added: Program management and leadership (n=42)
16. Added: Specific curriculum or teaching methods/technology (n=494)
17. Added: Child/classroom monitoring and assessment (n=99)
18. Added: Diversity skills: culture, language (n=28)
19. Added: Art, music, dance, expression (n=48)
20. Added: N/A: Responded about type of training, sponsorship or source of support, rather than content (n=19)
21. OTHER, including categories 10, 13, 15, 18, 19, and 20

-1. Don't know/Refused/No answer (n=149)
-2. Not applicable – didn't attend workshop/training to report on (n=482)

NOTES:

In order to minimize the risk of disclosure, categories 10, 13, 15, 18, 19 and 20 were grouped together into category 21 (Other). Total frequencies for each of the original response categories are also presented above (denoted by n=).

Please see the NSECE Workforce restricted-use data file for the more comprehensive variable and original response data.

WF_A9_PROFASSC (QT = WF_CAREER_PROFASSOC)

A9. Are you a member of a professional association focused on caring for children (such as the National Association for the Education of Young Children, the National Family Child Care Association, the National Institute on Out of School Time, a religiously identified child care organization, or a similar organization)?

1. Yes
2. No
3. Don't know/Refused/No answer

WF_A10_UNION (QT = WF_CAREER_UNION)

A10. Are you a member of a union (such as Service Employees International Union, American Federation of Teachers, American Federation of State, County and Municipal Employees (AFSCME) or the Teamsters)?

1. Yes
2. No
3. Don't know/Refused/No answer

WF_A11_REASON (QT = WF_CAREER_REASON)

A11. Which one of the following best describes the main reason that you work with young children? (*CODE ONE ONLY*).

1. It is my career or profession
2. It is a step towards a related career
3. It is my personal calling
4. It is a job with a paycheck
5. It is work I can do while my own children are young
6. It is a way to help children
7. It is a way to help parents
8. None of these reasons apply
9. Don't know/Refused/No answer

Section C. Activities

The next questions are about your activities with children.

WF_WORK_DAYS
(R) WF_C1_DAYSWORK

C1. Last week, how many days did you work at this program?

_____ Days
-1. DON'T KNOW/REFUSED/NO ANSWER

NOTES:

In order to minimize the risk of disclosure, responses that are more than 5 days a week have been grouped into a single category of "6 or more days a week".

Original response data is available in the NSECE WF restricted-use data file.

WF_C1_CURRICULUM (QT = WF_WORK_CRCLM)

C1A. Did you use a curriculum or prepared set of learning and play activities?

1. Yes ➔ ASK C1B
2. No ➔ GO TO C3
3. Don't know/Refused/No answer ➔ GO TO C3

WF_WORK_CRCLM_NAME
(R) WF_C1_NAMECURR_R

Variables affecting eligibility for this item: **WF_C1_CURRICULUM**

C1B. What is the name of the curriculum or approach used?

1. A curriculum we developed ourselves (n=1560)
2. Bank Street Developmental Interaction Approach (n=43)
3. Galileo (n=8)
4. Innovations Series Curriculum (n=5)
5. Learning Games
6. Montessori Infant/Toddler Curriculum (n=27)
7. Montessori Preschool Curriculum (n=10)
8. Opening the World of Learning (OWL) (n=27)
9. Preschool Paths (n=37)
10. Project Approach (n=79)
11. Reggio Emilia Approach (n=31)
12. Scholastic Early Childhood Program (n=18)

13. The Creative Curriculuma for Infants and Toddlers (n=19)
14. The Creative Curriculuma for Preschool (n=10)
15. The High/Scope Curriculum for Preschool (n=22)
16. The High/Scope Curriculum for Infants and Toddlers (n=39)
17. The Program for Infant/Toddler Caregivers (PITC) Curriculum (n=46)
18. Waldorf Approach (n=297)
19. Other (n=835)
20. None (n=200)
21. Don't know/Refused/No answer (n=131)
22. Added: Teaching Strategies Gold (n=54)
23. Added: Tools of the Mind (n=18)
24. Added: Montessori (Unspecified) (n=17)
25. Added: High/Scope (Unspecified) (n=136)
26. Added: Creative Curriculum (Unspecified) (n=271)
27. Added: Teaching Strategies (Unspecified) (n=30)
28. Added: Curricula dictated by host organization (n=85)
29. Added: Purchased/publicly available curricula (n=732)
30. Added: Activities/activity planning (n=24)
31. The Creative Curriculum for Infant/Toddler; Preschool; or Unspecified age group (collapses original categories 13-14, 26)
32. High/Scope for Infant/Toddler; Preschool; or Unspecified age group (collapses categories 15-16, 25)
33. Other (collapses categories 19, 2-12, 17, 22-24, 27-30)

-2. Not applicable - Did not use curriculum. (n=745)

NOTES:

A number of categories were collapsed into broader categories (31, 32, and 33) to minimize the risk of disclosure in the public-use data. Total frequencies for each category are presented above (denoted by n=).

Please see the WF restricted-use data file for the more comprehensive variable.

WF_C3_PLAN (QT = WF_WORK_PLAN)

C3. Do you plan or help plan the daily activities of the children in this classroom or group?

1. Yes → ASK C3A
2. No → SKIP TO C4
3. Don't know/Refused/No answer → SKIP TO C4

WF_C3_WHENPLAN (QT = WF_WORK_PLAN_WHEN)

Variables affecting eligibility for this item: **WF_C3_PLAN**

C3a. When do you plan daily activities?

C IDM Survey Questions

Survey Questions
Program Level
Revised 1/23/19

1. Have you received training and/or ongoing support on the following practices within the last year?

□ Leading regular, data-informed processes meant to help improve the quality of teaching and learning
□ Organizing and facilitating job-embedded professional learning (e.g. coaching/mentoring, peer learning groups, team lesson planning)
□ Ensuring coherent instructional guidance and systems to support teacher practice
□ Creating systems in support of family engagement practices
□ Including teachers and families in decision making
□ Addressing and ensuring equity
□ Building a trusting and supportive environment among all in the program community

2. **For the next set of questions, please consider how strongly you agree or disagree with the following statements:**

	Strongly Disagree	Disagree	Neither agree or disagree	Agree	Strongly Agree		Comments
	1	2	3	4	5		
a. Professional development opportunities are affordable for me							
b. Professional development opportunities are accessible for me (e.g. online, within your community, language diversity, etc.)							
c. Professional development opportunities are relevant to my job							
d. Professional development opportunities have improved my capacity to do my job							

RESEARCH BASED CURRICULUM

3. Please specify what curriculum(s) you use in your classroom(s)?

- Primary curriculum:
- Second curriculum (optional):
- Additional curriculum used (optional):
- If no curriculum, skip questions

4. Do you believe your curriculum(s) is/are evidence-based? *Answer for each curriculum*

a. Primary Curriculum: _____

- Yes
- No
- Don't Know
- Other, please explain: _____

b. Second Curriculum: _____

- Yes
- No
- Don't Know
- Other, please explain: _____

5. On a scale from 1 – 5, to what extent is/are your curriculum(s) aligned with state standards with 1 indicating not at all and 5 indicating completely? *Answer for each curriculum*

a. Primary Curriculum: _____

❑ 1 2 3 4 5

b. Second Curriculum: _____

❑ 1 2 3 4 5

6. To the best of your knowledge, is/are your curriculum(s) culturally responsive? *Answer for each curriculum*

a. Primary Curriculum: _____

- Yes
- No
- Don't Know
- Other, please explain: _____

b. Second Curriculum: _____

- Yes
- No

- Don't Know
- Other, please explain: _____

7. To the best of your knowledge, is/are your curriculum(s) linguistically responsive? *Answer for each curriculum*

a. Primary Curriculum: _____

- Yes
- No
- Don't Know
- Other, please explain: _____

b. Second Curriculum: _____

- Yes
- No
- Don't Know
- Other, please explain: _____

8. To the best of your knowledge, is/are your curriculum(s) supportive of individualized instruction for children with a range of abilities? *Answer for each curriculum*

a. Primary Curriculum: _____

- Yes
- No
- Don't Know
- Other, please explain: _____

b. Second Curriculum: _____

- Yes
- No
- Don't Know
- Other, please explain: _____

9. Do teachers and staff in your program receive formal training on implementing/using each of the curriculum?

- Yes
- No
- Don't Know

10. Do you have an established method of monitoring how effectively teachers are implementing curricula (e.g. use of a tool or assessment)?

- Yes
- No
- Don't Know
- Other, please explain: _____

11. If 'Yes', How many times a year is curriculum fidelity assessed?

- 1
- 2
- 3
- Other, please explain: _____Other, please explain: _____

12. If 'Yes', Does/Do your classroom (s) use data obtained from the curriculum monitoring tool for improvement?

- Yes
- No
- Don't Know
- Other, please explain: _____

CHILD ASSESSMENT

13. Does your program have formal processes for collecting child level data on cognitive and non-cognitive skill development? (e.g., through T.S. GOLD or a similar program)

- Yes
- No
- Don't Know
- Other, please explain: _____

14. If 'Yes', does your program(s) use one or more published tools to collect data on cognitive and non-cognitive skills of young children (e.g., T.S. GOLD)?

- Yes, the tool(s) is/are called: _____
- No

15. If 'Yes', are the assessments you use to evaluate children's cognitive and non-cognitive skills research based?

- Yes
- No

- Don't Know

Other, please explain: _____

16. If 'Yes', are the assessments you use comprehensive across all domains and development?

- Yes
- No
- Don't Know
- Other, please explain: _____

17. If 'Yes', are your assessments aligned to state learning and development standards?

- Yes
- No
- Don't Know
- Other, please explain: _____

18. How does your program use data from cognitive and non-cognitive assessments? Check all that apply:

- □ To inform instruction at the classroom level
- □ To inform professional development across the program
- □ To inform continuous improvement across the program
- □ N/A, my program does not use formative assessments
- □ Don't Know
- □ Other, please explain:_____

19. Are teachers in your program formally trained and certified in the formative assessments they use?

- Yes, they are trained by: _____
- No
- Don't Know

Other, please explain: _____

20. What supports for cognitive and non-cognitive child assessment implementation are available to teachers?

Check all that apply:

- Written materials
- In person training
- Online training
- Peer learning/teacher collaboration groups

Other, please explain: _____

HIGH QUALITY INSTRUCTION

21. Does your program implement all of the state's comprehensive learning and development standards for Pre-K?

- Yes, we do this by: _____
- No
- Don't Know

Other, please explain: _____

CONTINUOUS QUALITY IMPROVEMENT

22. Are programs required to conduct annual, or more frequent, program quality assessments for continuous quality improvement?

- Yes
- No
- Don't Know
- Other, please explain: _____

23. If 'Yes,' which types of data, if any, are programs required to use when conducting program quality assessments?

□ N/A

□ Student data (e.g., enrollment, attendance, assessments)

□ Student data analyzed by subgroup

□ Classroom observation / teaching effectiveness data

□ None of the above

□ Don't know

□ Other, please explain: _____

24. If 'Yes,' which processes of data analysis

25. If 'Yes,' which processes of data analysis, if any, are programs required to use to inform improvement efforts?

□ Leaders, staff, and other stakeholders collaborate to analyze data and create professional learning goals and plans

□ Programs set annual, or more frequent, goals and plans for improving teaching and learning, and actively monitor progress towards those goals

□ None of the above

□ Don’t know

□ Other, please explain:_____

D Survey Sampling Tutorial

1.2 Requirements of a Good Sample

In the movie "Magic Town," the public opinion researcher played by James Stewart discovered a town that had exactly the same characteristics as the whole United States: Grandview had exactly the same proportion of people who voted Republican, the same proportion of people under the poverty line, the same proportion of auto mechanics, and so on, as the United States taken as a whole. All that Stewart's character had to do was to interview the people of Grandview, and he would know what public opinion was in the United States.

A perfect sample would be like Grandview: a "scaled-down" version of the population, mirroring every characteristic of the whole population. Of course, no such perfect sample can exist for complicated populations (even if it did exist, we would not know it was a perfect sample without measuring the whole population). But a good sample will be **representative** in the sense that characteristics of interest in the population can be estimated from the sample with a known degree of accuracy.

Some definitions are needed to make the notion of a good sample more precise.

Observation unit An object on which a measurement is taken. This is the basic unit of observation, sometimes called an **element**. In studying human populations, observation units are often individuals.

Target population The complete collection of observations we want to study. Defining the target population is an important and often difficult part of the study. For example, in a political poll, should the target population be all adults eligible to vote? All registered voters? All persons who voted in the last election? The choice of target population will profoundly affect the statistics that result.

Sample A subset of a population.

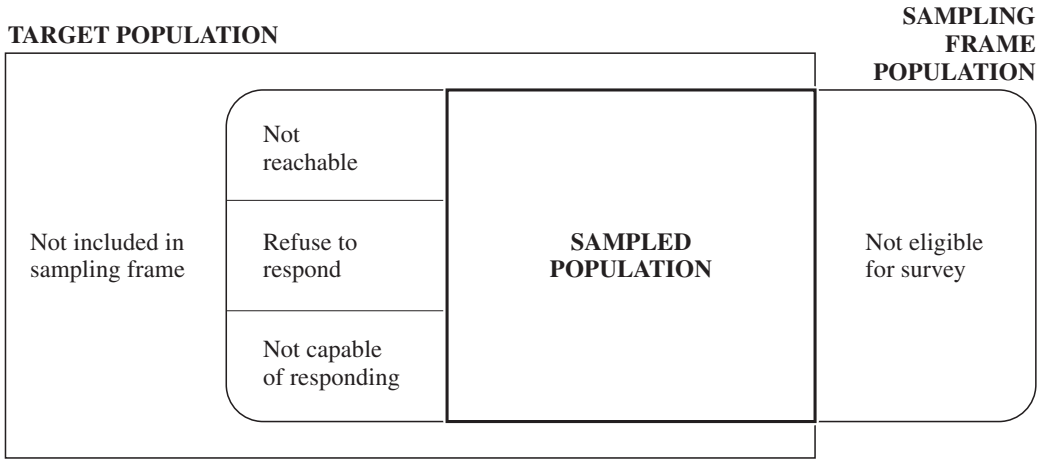
Sampled population The collection of all possible observation units that might have been chosen in a sample; the population from which the sample was taken.

Sampling unit A unit that can be selected for a sample. We may want to study individuals, but do not have a list of all individuals in the target population. Instead, households serve as the sampling units, and the observation units are the individuals living in the households.

Sampling frame A list, map, or other specification of sampling units in the population from which a sample may be selected. For a telephone survey, the sampling frame might be a list of all residential telephone numbers in the city. For a survey using in-person interviews, the sampling frame might be a list of all street addresses. For an agricultural survey, a sampling frame might be a list of all farms, or a map of areas containing farms.

In an ideal survey, the sampled population will be identical to the target population, but this ideal is rarely met exactly. In surveys of people, the sampled population is usually smaller than the target population: as illustrated in Figure 1.1, not all persons in the target population are included in the sampling frame, and a number of persons will not respond to the survey.

FIGURE 1.1
Target population and sampled population in a telephone survey of likely voters. Not all households have telephones, so a number of persons in the target population of likely voters will not be associated with a telephone number in the sampling frame. In some households with telephones, the residents are not registered to vote and hence are not eligible for the survey. Some eligible persons in the sampling frame population do not respond because they cannot be contacted, some refuse to respond to the survey, and some may be ill and incapable of responding.



In the Hite (1987) study, one characteristic of interest was the percentage of women who are harassed in their relationship. An individual woman was an element. The target population was all adult women in the United States. Hite's sampled population was women belonging to women's organizations who would return the questionnaire. Consequently, inferences can only be made to the sampled population, not to the population of all adult women in the United States.

The National Crime Victimization Survey (NCVS) is an ongoing survey to study victimization rates, administered by the U.S. Census Bureau and the Bureau of Justice Statistics. If the characteristic of interest is the total number of households in the United States that were victimized by crime last year, the elements are households, the target population consists of all households in the United States, and the sampled population consists of households in the sampling frame, constructed from U.S. Census information and building permits, that are "at home" and agree to answer questions.

The goal of the National Pesticide Survey, conducted by the U.S. Environmental Protection Agency, was to study pesticides and nitrate in drinking water wells nationwide. The target population was all community water systems and rural domestic wells in the United States. The sampled population was all community water systems (all are listed in the Federal Reporting Data System) and all identifiable domestic wells outside of government reservations that belonged to households willing to cooperate with the survey.

Public opinion polls are often taken to predict which candidate will win the next election. The target population is persons who will vote in the next election; the sampled population is often persons who can be reached by telephone and who are judged to be likely to vote in the next election. Few national polls in the

United States include persons in hospitals, dormitories, or jails; they, and persons without telephones, are not part of the sampling frame or of the sampled population.

1.3 Selection Bias

A good sample will be as free from selection bias as possible. **Selection bias** occurs when some part of the target population is not in the sampled population, or, more generally, when some population units are sampled at a different rate than intended by the investigator. If a survey designed to study household income omits transient persons, the estimates from the survey of the average or median household income are likely to be too large. A **sample of convenience** is often biased, since the units that are easiest to select or that are most likely to respond are usually not representative of the harder-to-select or nonresponding units. The following examples indicate some ways in which selection bias can occur.

- Using a sample selection procedure that, unknown to the investigators, depends on some characteristic associated with the properties of interest. For example, investigators took a convenience sample of adolescents to study how frequently adolescents talk to their parents and teachers about AIDS. But adolescents willing to talk to the investigators about AIDS are probably also more likely to talk to other authority figures about AIDS. The investigators, who simply averaged the amounts of time that adolescents in the sample said they spent talking with their parents and teachers, probably overestimated the amount of communication occurring between parents and adolescents in the population.
- Deliberately or purposively selecting a "representative" sample. If we want to estimate the average amount a shopper spends at the Mall of America in a shopping trip, and we sample shoppers who look like they have spent an "average" amount, we have deliberately selected a sample to confirm our prior opinion. This type of sample is sometimes called a **judgment sample**—the investigator uses his or her judgment to select the specific units to be included in the sample.
- Misspecifying the target population. For instance, all the polls in the 1994 Democratic gubernatorial primary election in Arizona predicted that candidate Eddie Basha would trail the front-runner in the polls by at least nine percentage points. In the election, Basha won 37% of the vote; the other two candidates won 35% and 28%, respectively. One problem is that many voters were undecided at the time the polls were taken. Another is that the target population for the polls was registered voters who had voted in previous primary elections and were interested in this one. In the primary election, however, Basha had heavy support in rural areas from demographic groups that had not voted before and hence were not targeted in the surveys.
- Failing to include all of the target population in the sampling frame, called **undercoverage**. The U.S. Behavioral Risk Factor Surveillance System survey, described at www.cdc.gov, illustrates some of the coverage problems that may occur in a household telephone survey. The target population for this survey on preventive

health practices and risk behaviors is adults aged 18 and older in the United States. Some undercoverage occurs because persons in institutions such as nursing homes or prisons are excluded. Additional undercoverage occurs because the survey is conducted by telephone. Some households do not have telephones and telephone coverage varies across states. Households in the southern part of the United States, minority households, and low-income households are less likely to have telephones, so those households are likely to be underrepresented in the sample because of the undercoverage. Households that have only a cellular telephone are also not included in the sampling frame at this writing.

- Including population units in the sampling frame that are not in the target population, called **overcoverage**. Overcoverage can occur when persons not in the target population are not screened out of the sample, or when data collectors are not given specific instructions on sample eligibility. The target population for a telephone survey on radio listening habits might be persons aged 18 and over, but some interviewers might include persons under age 18 when taking the sample, and children and teenagers may well listen to different radio stations than adults.
- Having multiplicity of listings in the sampling frame, without adjusting for the multiplicity in the analysis. In its simplest form, random digit dialing prescribes selecting a random sample of 10-digit numbers. Households with more than one telephone line then have a higher chance of being selected in the sample. This multiplicity can be compensated in the estimation (we'll discuss this in Section 6.5); if it is ignored, bias can result. One might expect households with more telephone lines to be larger or more affluent, so if no adjustment is made for those households having a higher probability of being selected for the sample, estimates of average income or household size may be too large.
- Substituting a convenient member of a population for a designated member who is not readily available. For example, if no one is at home in the designated household, a field representative might try next door. In a wildlife survey, the investigator might substitute an area next to a road for a less accessible area. In each case, the sampled units most likely differ on a number of characteristics from units not in the sample. The substituted household may be more likely to have a member who does not work outside of the house than the originally selected household. The area by the road may have fewer frogs than the area that is harder to reach.
- Failing to obtain responses from all of the chosen sample. **Nonresponse** distorts the results of many surveys, even surveys that are carefully designed to minimize other sources of selection bias. Often, nonrespondents differ critically from the respondents, but the extent of that difference is unknown unless you can later obtain information about the nonrespondents. Many surveys reported in newspapers or research journals have dismal response rates—in some, the response rate is as low as 10%. It is difficult to see how results can be generalized to the population when 90% of the targeted sample cannot be reached or refuses to participate.

  The Adolescent Health Database Survey was designed to obtain a representative sample of Minnesota junior and senior high school students in public schools (Remafedi et al., 1992). Overall, 49% of the school districts that were invited to

participate in the survey agreed to participate. The response rate varied with the size of the school district:

Type of School District	Participation Rate (%)
Urban	100
Metropolitan suburban	25
Nonmetropolitan with more than 2000 students	62
Nonmetropolitan with 1000–1999 students	27
Nonmetropolitan with 500–999 students	61
Nonmetropolitan with fewer than 500 students	53

In each of the school districts that participated, surveys were distributed to students and students' participation was voluntary. Of the 52,553 surveys distributed to students, 36,741 were completed and returned, resulting in a student response rate of 69%. The survey asked questions about health habits, religious affiliation, psychosocial status, and sexual orientation. It seems likely that responding and nonresponding school districts have different levels of health and activity. It seems even more likely that students who respond to the survey will, on average, have a different health profile than students who do not respond to the survey.

Many studies comparing respondents and nonrespondents have found differences in the two groups. In the Iowa Women's Health Study, 41,836 women responded to a mailed questionnaire in 1986. Bisgard et al. (1994) compared those respondents to the 55,323 nonrespondents by checking records in the State Health Registry; they found that the age-adjusted mortality rate and the cancer attack rate were significantly higher for the nonrespondents than for the respondents.

- Allowing the sample to consist entirely of volunteers. Such is the case in radio and television call-in polls, and in most online surveys. The statistics from such surveys cannot be trusted. At best, they are entertainment; at worst, they mislead, particularly when statistics from polls with self-selected respondents are cited in policy debates without any mention of their unscientific nature. CNN.com's daily QuickVote, which invites site visitors to vote on an issue of the day, carefully states that "This QuickVote is not scientific and reflects the opinions of only those Internet users who have chosen to participate. The results cannot be assumed to represent the opinions of Internet users in general, nor the public as a whole" (Cable News Network, 2002). Yet statistics from QuickVote and other online surveys are frequently quoted by independent research institutes, policy organizations, and scholarly journals. For example, Christian and Kinney (1999) cited a 1999 Internet poll on CNN.com, where 98% of the 17,000 visitors to a website linked to the science and technology reports voted "yes" to a question on whether the Hubble Space Telescope was worth the investment, as an indication of "a great improvement in public opinion." In fact, all that can be concluded from the Internet poll is that nearly 17,000 people who visited a website voted "yes" on the question; nothing can be inferred about the rest of the population without making heroic assumptions. Some individuals or organizations may respond multiple times to a voluntary survey, and a determined organization may skew the results.

EXAMPLE 1.1 Many surveys have more than one of these problems. *The Literary Digest* (1932, 1936a, b, c) began taking polls to forecast the outcome of the U.S. presidential election in 1912, and their polls attained a reputation for accuracy because they forecast the correct winner in every election between 1912 and 1932. In 1932, for example, the poll predicted that Roosevelt would receive 56% of the popular vote and 474 votes in the Electoral College; in the actual election, Roosevelt received 58% of the popular vote and 472 votes in the Electoral College.

With such a strong record of accuracy, it is not surprising that the editors of *The Literary Digest* had a great deal of confidence in their polling methods by 1936. Launching the 1936 poll, they said:

> The Poll represents thirty years' constant evolution and perfection. Based on the "commercial sampling" methods used for more than a century by publishing houses to push book sales, the present mailing list is drawn from every telephone book in the United States, from the rosters of clubs and associations, from city directories, lists of registered voters, classified mail-order and occupational data. (1936a, p. 3)

On October 31, the poll predicted that Republican Alf Landon would receive 55% of the popular vote, compared with 41% for President Roosevelt. The article "Landon, 1,293,669; Roosevelt, 972,897: Final Returns in The Digest's Poll of Ten Million Voters" contained the statement "We make no claim to infallibility. We did not coin the phrase 'uncanny accuracy' which has been so freely applied to our Polls" (1936b). It is a good thing they made no claim to infallibility: In the election, Roosevelt received 61% of the vote; Landon, 37%.

What went wrong? One problem may have been undercoverage in the sampling frame, which relied heavily on telephone directories and automobile registration lists—the frame was used for advertising purposes, as well as for the poll. Households with a telephone or automobile in 1936 were generally more affluent than other households, and opinion of Roosevelt's economic policies was generally related to the economic class of the respondent. But sampling frame bias does not explain all the discrepancy. Postmortem analyses of the poll by Squire (1988) and Calahan (1989) indicate that even persons with both a car and a telephone tended to favor Roosevelt, though not to the degree that persons with neither car nor telephone supported him.

The low response rate to the survey was likely the source of much of the error. *Ten million* questionnaires were mailed out, and 2.3 million were returned—an enormous sample, but a response rate of less than 25%. In Allentown, Pennsylvania, for example, the survey was mailed to every registered voter, but the survey results for Allentown were still incorrect because only one-third of the ballots were returned. Squire (1988) reports that persons supporting Landon were much more likely to have returned the survey; in fact, many Roosevelt supporters did not even remember receiving a survey even though they were on the mailing list.

One lesson to be learned from *The Literary Digest* poll is that the sheer size of a sample is no guarantee of its accuracy. The *Digest* editors became complacent because they sent out questionnaires to more than one quarter of all registered voters and obtained a huge sample of 2.3 million people. But large unrepresentative samples can perform as badly as small unrepresentative samples. A large unrepresentative sample may do more damage than a small one because many people think that large

samples are always better than small ones. The design of the survey is far more important than the absolute size of the sample. ■

What good are samples with selection bias? We prefer to have samples with no selection bias, that serve as a microcosm of the population. When the primary interest is in estimating the total number of victims of violent crime in the United States, or the percentage of likely voters in the United Kingdom who intend to vote for the Labour Party in the next election, serious selection bias can cause the sample estimates to be invalid.

Purposive or judgment samples can provide valuable information, though, particularly in the early stages of an investigation. Teichman et al. (1993) took soil samples along Interstate 880 in Alameda County, California, to determine the amount of lead in yards of homes and in parks close to the freeway. In taking the samples, they concentrated on areas where they thought children were likely to play and areas where soil might easily be tracked into homes. The purposive sampling scheme worked well for justifying the conclusion of the study, that "lead contamination of urban soil in the east bay area of the San Francisco metropolitan area is high and exceeds hazardous waste levels at many sites." A sampling scheme that avoided selection bias would be needed for this study if the investigators wanted to generalize the estimated percentage of contaminated sites to the entire area.

1.4 Measurement Error

A good sample has accurate responses to the items of interest. When a response in the survey differs from the true value, **measurement error** has occurred. **Measurement bias** occurs when the response has a tendency to differ from the true value in one direction. As with selection bias, measurement error and bias must be considered and minimized in the design stage of the survey; no amount of statistical analysis will disclose that the scale erroneously added 5 kilograms to the weight of every person in the health survey.

Measurement error is a concern in all surveys and can be insidious. In many surveys of vegetation, for example, areas to be sampled are divided into smaller plots. A sample of plots is selected, and the number of plants in each plot is recorded. When a plant is near the boundary of the region, the field researcher needs to decide whether to include the plant in the tally. A person who includes all plants near or on the boundary in the count is likely to produce an estimate of the total number of plants in the area that is too high because some plants may be counted twice. Duce et al. (1972) report concentrations of trace metals, lipids, and chlorinated hydrocarbons in the top 100 micrometers of Narragansett Bay that are 1.5 to 50 times as great as those in the water 20 cm below the surface. If studying the transport of pollutants from coastal waters to the deeper waters of the ocean, a sampling scheme that ignores this boundary effect may underestimate the amount transported.

Sometimes measurement bias is unavoidable. In the North American Breeding Bird Survey, observers stop every one-half mile on designated routes and count all birds heard or seen during a 3-minute period within a quarter-mile radius

(Sauer et al., 1997). The count of birds for that point is almost always an underestimate of the number of birds in the area; statistical models may possibly be used to adjust for the measurement bias. If data are collected with the same procedure and with similarly skilled observers from year to year, the survey can be used to estimate trends in the population of different species—the biases from different years are expected to be similar, and may cancel when year-to-year differences are calculated.

Obtaining accurate responses is challenging in all types of surveys, but particularly so in surveys of people:

- People sometimes do not tell the truth. In an agricultural survey, farmers in an area with food-aid programs may underreport crop yields, hoping for more food aid. Obtaining truthful responses is a particular challenge in surveys involving sensitive subject matter, such as surveys about drug use.
- People do not always understand the questions. Many persons in the United States were shocked by the results of a 1993 Roper poll reporting that 25% of Americans did not believe the Holocaust really happened. When the double-negative structure of the question was eliminated, and the question reworded, only 1% thought it was "possible … the Nazi extermination of the Jews never happened."
- People forget. One problem faced in the design of the NCVS is that of **telescoping**: Persons are asked about experiences as a crime victim in the last six months, but some include victimizations that occurred more than six months ago.
- People give different answers to different interviewers. Schuman and Converse (1971) employed white and black interviewers to interview black residents of Detroit. In response to the question "Do you personally feel that you can trust most white people, some white people, or none at all?" 35% of the respondents interviewed by a white person said they could trust most white people. The percentage was 7% for those interviewed by a black person.
- People may say what they think an interviewer wants to hear or what they think will impress the interviewer. In experiments done with questions beginning "Do you agree or disagree with the following statement" it has been found that a subset of the population tends to agree with any statement regardless of its content. Lenski and Leggett (1960) found that about one-tenth of their sample agreed with both of the following statements:

  > It is hardly fair to bring children into the world, the way things look for the future.
  >
  > Children born today have a wonderful future to look forward to.

  Some responses are perceived as being more socially desirable than others, so that persons may overreport behaviors such as exercising and donating to charities, and underreport behaviors such as smoking or drinking.
- A particular interviewer may affect the accuracy of the response, by misreading questions, recording responses inaccurately, or antagonizing the respondent. In a survey on abortion, a poorly trained interviewer with strong feelings about abortion may encourage the respondent to provide one answer rather than another. In extreme cases, an interviewer may change the answers given by the respondent, or simply make up data and not contact the respondent at all.

- Certain words mean different things to different people. A simple question such as "Do you own a car?" may be answered yes or no depending on the respondent's interpretation of "you" (does it refer to just the individual, or to the household?), "own" (does it count as ownership if you are making payments to a finance company?), or "car" (are pickup trucks included?).
- Question wording and question order have a large effect on the responses obtained. Two surveys were taken in late 1993/early 1994 about Elvis Presley. One survey asked, "In the past few years, there have been a lot of rumors and stories about whether Elvis Presley is really dead. How do you feel about this? Do you think there is any possibility that these rumors are true and that Elvis Presley is still alive, or don't you think so?" The other survey asked, "A recent television show examined various theories about Elvis Presley's death. Do you think it is possible that Elvis is alive or not?" Eight percent of the respondents to the first question said it is possible that Elvis is still alive; 16% of respondents to the second question said it is possible that Elvis is still alive.

Excellent discussions of these problems can be found in Groves et al. (2009) and Tourangeau et al. (2000). In some cases, accuracy can be increased by careful questionnaire design.

1.5 Questionnaire Design

This section gives a very brief introduction to writing and testing questions. It provides some general guidelines and examples, but if you are writing a questionnaire, you should consult one of the more comprehensive references on questionnaire design listed at the end of this chapter.

The most important step in writing a questionnaire is to decide what you want to find out. Write down the goals of your survey, and be precise. "I want to learn something about the homeless" won't do. Instead, you should write down specific questions, such as "What percentage of persons using homeless shelters in Chicago between January and March 1996 are under 16 years old?" Then, write or select questions that will elicit accurate answers to the research questions, and that will encourage persons in the sample to respond to the questions.

- *Always test your questions before taking the survey.* Ideally, the questions would be tested on a small sample of members of the target population. Try different versions for the questions, and ask respondents in your pretest how they interpret the questions.

  The NCVS was tested for several years before it was conducted on a national scale (Lehnen and Skogan, 1981). The pretests were used to help decide on a recall period (it was decided to ask respondents about victimizations that had occurred in the previous six months), test interviewing procedures and questions, and compare information from selected interviews with information found in the police report about the victimization. As a result of the pretests, some of the long and repetitious questions were shortened and more specific wording introduced.

The questionnaire was revised in 1985 and again in 1991 to make use of recent research in cognitive psychology and to include topics, such as victim and bystander behavior, that were not found in the earlier versions. All revisions are tested extensively in the field before being used (Taylor, 1989). In the past, for example, the NCVS has been criticized for underreporting the crime of rape; when the questionnaire was designed in the early 1970s, there was worry that asking about rape directly would be perceived as insensitive and embarrassing, and would provoke congressional outrage. The original NCVS questionnaire asked a series of specific questions intended to prompt the memory of respondents. These included questions such as "Did anyone take something directly from you by using force, such as by a stickup, mugging or threat?" The last question in the violent crime screening section of the questionnaire was "Did anyone try to attack you in some other way?" If the respondent mentioned in response that he or she was raped, then a rape was reported. Not surprisingly, the victimization rate for rape reported for the 1990 and earlier NCVS is very low: It is reported that about 1 per 1000 females aged 12 and older were raped in 1990. The current version of the NCVS questionnaire asks about rape directly.

You will not necessarily catch misinterpretations of questions by trying them out on friends or colleagues; your friends and colleagues may have backgrounds similar to yours, and may not have the same understanding of words as persons in your target population. Belson (1981) demonstrated that each of 29 questions about television viewing was misinterpreted by some respondents. The question "Do you think that the television news programmes are impartial about politics?" was tested on 56 people. Of these, 13 interpreted the question as intended, 18 respondents narrowed the term *news programmes* to mean "news bulletins," 21 narrowed it to "political programmes," and 1 interpreted it as "newspapers." Only 25 persons interpreted "impartial" as intended; 5 inferred the opposite meaning, "partial"; 11, as "giving too much or too little attention to"; and the others were simply unfamiliar with the word. Suessbrick et al. (2000) found that the concepts in a seemingly clear question such as "Have you smoked at least 100 cigarettes in your entire life?" were commonly interpreted in a different way than the authors intended: Some respondents included marijuana cigarettes or cigars, while others excluded cigarettes that were only partially smoked or hand-rolled cigarettes.

- *Keep it simple and clear.* Questions that seem clear to you may not be clear to someone listening to the whole question over the telephone, or to a person with a different native language. Belson (1981, p. 240) tested the question "What proportion of your evening viewing time do you spend watching news programmes?" on 53 people. Only 14 people correctly interpreted the word "proportion" as "percentage," "part," or "fraction." Others interpreted it as "how long do you watch" or "which news programmes do you watch."
- *Use specific questions instead of general ones, if possible.* Strunk and White advised writers to "Prefer the specific to the general, the definite to the vague, the concrete to the abstract" (1959, p. 15). Good questions result from good writing.

  Instead of asking "Did anyone attack you in the last six months," the NCVS asks a series of specific questions detailing how one might be attacked. The NCVS question is "Has anyone attacked or threatened you in any of these ways: (a) With

any weapon, for instance, a gun or knife, (b) With anything like a baseball bat, frying pan, scissors, or stick …."

- *Relate your questions to the concept of interest.* This seems obvious but is forgotten or ignored in many surveys. In some disciplines, a standard set of questions has been developed and tested, and these are then used by subsequent researchers. Often, use of a common survey instrument allows results from different studies to be compared. In some cases, however, the standard questions are inappropriate for addressing the research hypotheses.

  Pincus (1993) criticizes early research that concluded that persons with arthritis were more likely to have psychological problems than persons without arthritis. In those studies, persons with arthritis were given the Minnesota Multiphasic Personality Inventory, a test of 566 true/false questions commonly used in psychological research. Patients with rheumatoid arthritis tended to have high scores on the scales of hypochondriasis, depression, and hysteria. Part of the reason they scored highly on those scales is clear when the actual questions are examined. A person with arthritis can truthfully answer false to questions such as "I am about as able to work as I ever was," "I am in just as good physical health as most of my friends," and "I have few or no pains" without being either hysterical or a hypochondriac.

- *Decide whether to use open or closed questions.* An **open question** allows respondents to form their own response categories; in a **closed question** (multiple choice), the respondent chooses from a set of categories read or displayed. Each has advantages. A closed question may prompt the respondent to remember responses that might otherwise be forgotten, and is in accordance with the principle that specific questions are better than general ones. If the subject matter has been thoroughly pretested and responses of interest are known, a well-written closed question will usually elicit more accurate responses, as in the NCVS question "Has anyone attacked or threatened you with anything like a baseball bat, frying pan, scissors, or stick?" If the survey is exploratory or questions are sensitive, though, it is often better to use an open question: Bradburn and Sudman (1979) note that respondents reported higher frequency of drinking alcoholic beverages when asked an open question than a closed question with categories "never" through "daily."

  Schuman and Scott (1987) conclude that, depending on the context, either open or closed questions can limit the types of responses received. In one experiment, the most common responses to the open question "What do you think is the most important problem facing this country today?" were "unemployment" (17%) and "general economic problems" (17%). The closed version asked, "Which of the following do you think is the most important problem facing this country today—the energy shortage, the quality of public schools, legalized abortion, or pollution—or if you prefer, you may name a different problem as most important"; 32% or respondents chose "the quality of public schools." In this case, the limited options in the closed question guided respondents to one of the listed responses. In another experiment, Schuman and Scott (1987) asked respondents to name one or two of the most important national world events or changes during the last 50 years. Persons asked the open question most frequently gave responses such

as World War II or the Vietnam War; they typically did not mention events such as the invention of the computer, which was the most prevalent response to the closed question including this option.

If using a closed question, always have an "other" category. In one study of sexual activity among adolescents, adolescents were asked from whom they felt the most pressure to have sex. Categories for the closed question were "friends of same sex," "boyfriend/girlfriend," "friends of opposite sex," "TV or radio," "don't feel pressure," and "other." The response "parents" or "father" was written in by a number of the adolescent respondents, a response that had not been anticipated by the researchers.

- *Report the actual question asked.* Public opinion is complex, and you inevitably leave a distorted impression of it when you compress the results of your careful research into a summary statement "*x*% of Americans favor affirmative action."

  The results of three surveys in Spring 1995, all purportedly about affirmative action, emphasize the importance of reporting the question. A *Newsweek* poll asked "Should there be special consideration for each of the following groups to increase their opportunities for getting into college and getting jobs or promotions?" and asked about these groups: blacks, women, Hispanics, Asians, and Native Americans. The poll found that 62% of blacks but only 25% of whites answered "yes" to the question about blacks. A *USA Today*–CNN–Gallup poll asked the question "What is your opinion on affirmative action programs for women and minorities: do you favor them or oppose them?" and reported that 55% of respondents favored such programs. A Harris poll asking "Would you favor or oppose a law limiting affirmative action programs in your state?" reported 51% of respondents favoring such a law. These questions are clearly addressing different concepts because the differences in percentages obtained are too great to be ascribed to the different samples of people taken by the three organizations. Yet all three polls' results were described in newspapers in terms of percentages of persons who support affirmative action.

- *Avoid questions that prompt or motivate the respondent to say what you would like to hear.* These are often called **leading,** or **loaded, questions.** The May 17, 1994 issue of *The Wall Street Journal* reported the following question asked by the Gallup Organization in a survey commissioned by the American Paper Institute: "It is estimated that disposable diapers account for less than 2 percent of the trash in today's landfills. In contrast, beverage containers, third-class mail and yard waste are estimated to account for about 21 percent of trash in landfills. Given this, in your opinion, would it be fair to tax or ban disposable diapers?"

- *Consider the social desirability of responses to questions, and write questions that elicit honest responses.* Abelson et al. (1992) review several studies that find many people say they voted in the last election when they actually did not vote. They argue that voting is a socially desirable behavior, and many respondents do not want to admit that they did not vote; respondents need to be prompted to report their actual behavior.

- *Avoid double negatives.* Double negatives needlessly confuse the respondent. A question such as "Do you favor or oppose not allowing drivers to use cell phones

while driving?" might elicit either "favor" or "oppose" from a respondent who thinks persons should not use cell phones while driving.

- *Use forced-choice, rather than agree/disagree questions.* As noted earlier, some persons will agree with almost any statement. Schuman and Presser (1981, p. 223) report the following differences from an experiment comparing agree/disagree with forced-choice versions:

  Q1: Do you agree or disagree with this statement: Most men are better suited emotionally for politics than are most women.

  Q2: Would you say that most men are better suited emotionally for politics than are most women, that men and women are equally suited, or that women are better suited than men in this area?

	Years of schooling		
	0–11	12	13+
Q1: Percent "agree"	57	44	39
Q2: Percent "men better suited"	33	38	28

- *Ask only one concept per question.* In particular, avoid what are sometimes called **double-barreled** questions, so named because if one barrel of the shotgun does not get you, the other one will.

  The question "Do you agree with Bill Clinton's $50 billion bailout of Mexico?" appeared on a survey distributed by a member of the U.S. House of Representatives to his constituents. The question is really confusing two opinions of the respondent: the opinion of Bill Clinton, and the opinion of the Mexico policy. Disapproval of either one will lead to a "disagree" answer to the question. Note also the loaded content of the word *bailout,* which will almost certainly elicit more negative responses than the term *aid package* would.

- *Pay attention to question order effects.* If you ask more than one question on a topic, it is usually (but not always) better to ask the more general question first and follow it by the specific questions. McFarland (1981) conducted an experiment in which half of the respondents were given general questions (for example, "How interested would you say you are in religion: very interested, somewhat interested, or not very interested") first, followed by specific questions on the subject ("Did you, yourself, happen to attend church in the last seven days?"); the other half were asked the specific questions first and then asked the general questions. When the general question was asked first, 56% reported that they were "very interested in religion"; the percentage rose to 64% when the specific question was asked first.

  Serdula et al. (1995) found that in the years in which a respondent of a health survey was asked to report his or her weight and then immediately asked "Are you trying to lose weight?" 28.8% of men and 48.0% of women reported that they were trying to lose weight. When "Are you trying to lose weight?" was asked in the middle of the survey and the self-report question on weight at the end of the survey, 26.5% of the men and 40.9% of the women reported that they were trying to lose weight. The authors speculate that respondents who are reminded of their weight status may overreport trying to lose weight.

The 2000 U.S. Census had separate questions for race (with categories white, black, American Indian or Alaskan Native, Asian Indian, Chinese, and Filipino, among others) and ethnicity (with categories non-Hispanic, Mexican-American, Puerto Rican, Cuban, and other Hispanic). These are considered separate classifications; an individual can, for instance, be white Cuban or black non-Hispanic. The Census Bureau has done a great deal of experimental research to determine effects of alternate wordings and orderings of these questions on responses (Bates et al., 1995). Martin et al. (2005) report results of experiments comparing the questions used for the 1990 Census with those used for the 2000 Census. In 1990, race was question 4 and ethnicity was question 7; in 2000, ethnicity was question 7 and race was question 8. When the question on race occurred first, as in the 1990 Census, some Hispanic respondents looked for a Hispanic category, did not find it, and checked the "Other Race" category. After answering the race question, some persons skipped the ethnicity question so that there was substantial nonresponse on the ethnicity question. The reversed question order and other changes in the 2000 Census led to less missing data on both the race and ethnicity questions.

1.6 Sampling and Nonsampling Errors

Most opinion polls that you see report a "margin of error." Many merely say that the margin of error is 3 percentage points. Others give more detail, as in this excerpt from a *New York Times* poll: "In theory, in 19 cases out of 20 the results based on such samples will differ by no more than three percentage points in either direction from what would have been obtained by interviewing all Americans." The margin of error given in polls is an expression of **sampling error**, the error that results from taking one sample instead of examining the whole population. If we took a different sample, we would most likely obtain a different sample percentage of persons who visited the public library last week. Sampling errors are usually reported in probabilistic terms. We discuss the calculation of sampling errors for different survey designs in Chapters 2 through 7.

Selection bias and measurement error are examples of **nonsampling errors**, which are any errors that cannot be attributed to the sample-to-sample variability. In many surveys, the sampling error that is reported for the survey may be negligible compared to the nonsampling errors; you often see surveys with a 30% response rate proudly proclaiming their 3% margin of error, while ignoring the tremendous selection bias in their results.

The goal of this chapter was to sensitize you to various forms of selection bias and inaccurate responses. We can reduce some forms of selection bias by using probability sampling methods, as described in the next chapter. Accurate responses can often be achieved through careful design and testing of the survey instrument, thorough training of interviewers, and pretesting the survey. We shall return to nonsampling errors in Chapter 8, where we discuss methods that have been proposed for trying to reduce nonresponse error after the survey has been collected (sneak preview: none of

E CLASS Scores Across our Focus States

Report on Head Start CLASS Data for Fiscal Years 2012-2015 published by the Administration of Children and Families (ACF) shows there is not much of a difference in terms of CLASS ratings between various states as Figure E shows (these are Head Start ratings rather than state Pre-K ratings, however, as the latter are not available to us at the moment, we will use the former). This could justify using the same set of input parameters for the power analysis across our focus states.

Figure 14: ACF's Study of CLASS Scores Across Grantees in the US

Average CLASS® Domain Scores by State
Combined Across Fiscal Years 2012 - 2015

State	Number of grantees	Emotional Support	Classroom Organization	Instructional Support
Massachusetts	27	6.1	5.8	3.0
Michigan	39	6.1	5.7	3.0
Minnesota	36	6.1	5.8	3.0
Mississippi	14	5.8	5.5	2.6
Missouri	20	5.9	5.4	2.9
Montana	19	6.0	5.5	3.0
Nebraska	16	6.0	5.6	2.8
Nevada	5	6.1	6.0	2.8
New Hampshire	5	6.0	5.4	3.0
New Jersey	17	6.0	5.7	3.0
New Mexico	25	6.0	5.5	2.5
New York	78	6.0	5.7	2.9
North Carolina	53	6.0	5.7	2.8
North Dakota	14	6.0	5.7	2.8
Northern Mariana Islands	1	6.4	6.0	2.4
Ohio	50	6.0	5.6	2.9
Oklahoma	31	5.8	5.5	2.6
Oregon	26	6.1	5.8	2.9
Pennsylvania	50	6.0	5.7	3.0
Puerto Rico	17	6.0	5.6	2.7
Rhode Island	9	6.2	6.0	3.3
South Carolina	15	5.7	5.4	2.5
South Dakota	13	5.8	5.4	2.9
Tennessee	19	6.0	5.8	3.1
Texas	81	5.9	5.6	2.8
Utah	8	6.1	5.9	3.3
Vermont	8	6.4	6.2	3.7
Virgin Islands	1	5.7	5.3	2.2
Virginia	45	5.9	5.7	2.9
Washington	49	6.1	5.8	3.0
West Virginia	22	6.0	5.6	2.8
Wisconsin	38	6.0	5.7	2.8
Wyoming	10	6.2	5.8	2.9