EDF 7463 Analysis of Survey, Record, and other Qualitative Data
Lecture Notes

Introduction to Survey and Qualitative Research

I. Research Study Designs: Different Kinds

EDF 7463 largely focuses on three of the eight types of research identified below.

A. Survey Research
B. Qualitative Research
C. Correlational Research
D. Experimental Research
E. Quasi-Experimental Research
F. Single subject Research: Experimental and Nonexperimental
G. Observational Research
H. Historical Research

A. Purposes of Surveys

1. **Public opinion polls** are descriptive surveys that are used to determine how different groups of people feel about political, social, educational or economic issues.

2. **Developmental surveys** are concerned primarily with variables that differentiate people at different levels of age, growth, or maturation along a number of dimensions such as intellectual, physical, emotional, or social development.

3. **Follow-up surveys** are conducted to determine the status of a group after some period of time.

B. Classification of Surveys

1. **Cross sectional surveys:** involve the collection of data from people on a single occasion.

2. **Longitudinal surveys:** involve collecting data multiple times to measure change over time. Developmental surveys tend to be longitudinal in nature.

   a. **Panel studies**: involve surveying the same group of people over time as they grow and change. The same participants involved in the study are surveyed time and time again until the conclusion of the study.

b. **Trend studies**: involve surveying multiple groups of people at a particular stage in their life. Each group of people included in the study differs from the other groups in the study via the time at which they are surveyed. So that a trend may be detected over time, a different group of people may be surveyed every year for several years. What makes these groups similar is that they are all at the same developmental level; what makes them different is that they are surveyed at different times.

c. **Cohort studies**: involve surveying the same population of people over time as they grow and change. The trick here is that each time the survey is administered a different set of people from the same population is participating in the study. In other words, each sample from this population is different.

d. **Follow-up studies**: Similar to a panel study, though it is undertaken only after (sometimes long after) the panel study has been completed.

C. Three Data Collection Methods in Survey Research

1. Questionnaires, by Mail, E-mail or the Internet

a. Advantages
   - i. Inexpensive
   - ii. Can be confidential or anonymous
   - iii. Easy to score most items
   - iv. Standardized items and procedures

b. Disadvantages
   - v. Response rate may be small
   - vi. Cannot probe or explain items
   - vii. Only used by people who can read
   - viii. Possibility of response sets

2. Interviews

a. Advantages
  - i. Can probe and explain items
  - ii. Usually high return rate
  - iii. Can be recorded for later analysis
  - iv. Flexibility of use

b. Disadvantages
  - v. Time-consuming to use
  - vi. No anonymity
  - vii. Bias of the interviewer
  - viii. Complex scoring of unstructured items
  - ix. Training items

3. Telephones

a. Advantages
  - i. High response rate (as Dillman says, this is quickly changing)
  - ii. Quick data collection
  - iii. Can reach a wide range of locales and respondents

b. Disadvantages
  - iv. Requires phone numbers
  - v. Difficult to get in-depth data
  - vi. Requires training

D. Four Sources of Survey Error

1. Sampling error: The result of surveying only some, and not all, elements of the survey population
2. Coverage error: The result of not allowing all members of the survey population have on equal or known nonzero chance of being sampled for participation in the survey
3. Measurement Error: The result of poor question wording or questions being presented in such a way that inaccurate and interpretable answers are obtained
4. Nonresponse error: The result of people who respond to a survey being different from sampled individuals who do not respond, in a way relevant to the study

Survey Questionnaire Development

II. **The Blueprint**: Planning the Development of a Survey Questionnaire

1. Blueprint Table: A table used to define the domain for measures, surveys, performance tasks, etc. This table serves as an organizer that frames the major content categories and skills to be assessed. The proportion of the tasks or items that will be included on the instrument or overall performance should correspond roughly with how important the domain is relative to other domains. One way of gauging the importance of a domain is by considering how much time you spend on a topic during instruction. Some of these examples were developed for measure, but not surveys. Nevertheless, they are useful examples of different Blueprints.

Example #1

Freshman Survey: Blueprint Table

Content Base Category	14
Get Real	2
Let's eat	2
What's up with UREC	1
Fitness Scavenger Hunt	1
Roommate Contract Hall meeting	1
Four stages of Drinking	1
Multicultural Services	1
Bicycle Registration	2
Rape is NOT Sex	1
Faith MAPS (Religious backgrounds)	1

Example #2

International Student Survey: Blueprint Table

Content Base Category	
Student Classification (freshman, transfer, graduate)	1
Visa Classification (F-1, J-1, G-4, E-2, L-2, others)	1
Goals for International Student Learning	2
American Culture	1
Foreign Relationship abilities (in the United States)	4
Visa Status Laws	5
Issues concerning Academic Life and Student Learning (employment and housing)	3

Example #3

Multicultural Services: Blueprint Table

Content Base Category	Number of items
Discrimination/Racism	7
Cultural differences	7
Lack of Visible Culture	5
Under Representation	5
Degree of “Fit”	5
Leadership skills	5
Study skills	5
Role Models	5
Language Barriers	5
Social Interactions	5
Recruitment of Students	4
Retention of Students	4
Financial Concerns	4
Family Issues	4
Hate Crimes	4

Example #4

Sportsmanship Instrument: Blueprint Table

Content Base Category	34	Knowledge	Application	Evaluation
Define sportsmanship	16	3	5	8
Relate sportsmanship to game situations	5	5		
Appropriately model sportsmanship (team captains)	7		5	2
Articulate the value of sportsmanship	6	6		

2. Eight Criteria for Assessing Each Survey Question Constructed.

    a. Does the question require an answer?
    b. To what extent do survey recipients already have an accurate, ready-made answer for the question they are being asked to report?
    c. Can people accurately recall and report past behaviors?
    d. Is the respondent willing to reveal the requested information?
    e. Will the respondent feel motivated to answer each question?
    f. Is the respondent's understanding of response categories likely to be influenced by more than words?
    g. Is the survey information being collected by more than one mode?
    h. Is changing a question acceptable to the survey sponsor?

3. Choosing the Most Appropriate Question Structure.

    a. Open-ended Questions

    b. Close-ended Questions
        i. Close ended Questions with Ordered Response Categories.

        ii. Close ended Questions with Unordered Response Categories.

        iii. Partially Close-ended Questions with Unordered Response Categories.

4. Principles for Writing Survey Questions.

Principle 2.1 Choose simple over specialized words. Use vocabulary that can be understood by the respondents.

Principle 2.2 Choose as few words as possible to pose the question. Statements should be short, rarely exceeding 20 words.

Principle 2.3 Use complete sentences to ask questions. Each statement should be a proper grammatical sentence.

Principle 2.4 Avoid vague quantifiers when more precise estimates can be obtained.

Principle 2.5 Avoid specificity that exceeds the respondent's potential for having an accurate, ready-made answer.

Principle 2.6 Use equal numbers or positive and negative categories for scalar questions. In other words, try to have an almost equal number of statements expressing positive and negative feelings.

Principle 2.7 Distinguish undecided from neutral by placement at the end of the scale.

Principle 2.8 Avoid bias from unequal comparisons.

Principle 2.9 State both sides of attitude scales in the question stems.

Principle 2.10 Eliminate check-all-that-apply question formats to reduce primacy effects.

Principle 2.11 Develop response categories that are mutually exclusive.

Principle 2.12 Use cognitive design techniques to improve recall.

Principle 2.13 Provide appropriate time referents.

Principle 2.14 Be sure each question is technically accurate.

Principle 2.15 Choose wordings that allow essential comparisons to be made with previously collected data.

Principle 2.16 Avoid asking respondents to say yes in order to mean no.

Principle 2.17 Avoid double-barreled questions.

Principle 2.18 Soften the impact of potentially objectionable questions.

Principle 2.19 Avoid asking respondents to make unnecessary calculations.

Principle 2.20 Whenever possible, statements should be in simple sentences, rather than complex or compound sentences.

Principle 2.21 Do not use statements that are factual or capable of being interpreted as factual.

Principle 2.22 Avoid statements that can have more than one interpretation.

Principle 2.23 Avoid statements that are likely to be endorsed by almost everyone or almost no one.

Principle 2.24 Avoid statements containing universals such as all, always, none and never because they often introduce ambiguity.

Principle 2.25 Avoid using indefinite qualifiers such as only, just, merely, many, few, or seldom.

Principle 2.26 Avoid statements that contain “if” or “because” clauses.

Principle 2.27 Avoid use of negatives (e.g., not, none, never)

III. **The Assembly**: Developing the Survey Questionnaire

A. Constructing the Questionnaire in Three Steps.

Step 1. Define a desired navigational path for reading all information presented on each page of the questionnaire.

Principle 3.1) Write each question in a way that minimizes the need to re-read portions in order to comprehend the response task.

Principle 3.2) Place instructions exactly where that information is needed and not at the beginning of the questionnaire.

Principle 3.3) Place items with the same response categories into an item-in-a-series format, but do it carefully.

Principle 3.4) Ask one question at a time.

Principle 3.5) Minimize the use of matrices.

Step 2. Creating visual navigational guides that will assist respondents in adhering to the prescribed navigational path and correctly interpreting the written information. Before enumerating Principles 3.6 – 3.26, it is important to define six visual elements that contribute to the quality of a survey's construction.

Six visual elements of words and other symbols should be considered before the principles in Step 2 are presented

Visual Element 1. Increase the size of written elements to attract attention.
Visual Element 2. Increase the brightness or color (shadings) of visual elements to attract attention and establish appropriate groupings.
Visual Element 3. Use spacing to identify appropriate groupings of visual elements.
Visual Element 4. Use similarity to identify appropriate groupings of visual elements.
Visual Element 5. Maintain a consistent figure/ground format to make the response task easier.
Visual Element 6. Maintain simplicity, regularity, and symmetry to make the response task easier.

Principle 3.6) Begin by asking questions in the upper left quadrant; place any information not needed by the respondent in the lower right quadrant.

Principle 3.7) Use the largest and/or brightest measure symbols to identify the starting point on each page.

Principle 3.8) Identify the beginning of each succeeding question in a consistent way.

Principle 3.9) Number questions consecutively and simply, from beginning to end.

Principle 3.10) Use a consistent figure/ground format to encourage the reading of all words.

Principle 3.11) Limit the use of reverse print to section headings and/or question numbers.

Principle 3.12) Place more blank space between questions than between the subcomponents of questions.

Principle 3.13) Use dark print for questions and light print for answer choices.

Principle 3.14) Place special instructions inside of question numbers and not as freestanding entities.

Principle 3.15) Optional or occasionally needed instructions should be separated from the question's statement by font or symbol variations.

Principle 3.16) Do not place instructions in a separate instruction book or in a separate section of the questionnaire.

Principle 3.17) Use of lightly shaded colors as background fields on which to write all questions provides an effective navigational guide to respondents.

Principle 3.18) When shaded background fields are used identification of all answer spaces in white helps reduce item nonresponse.

Principle 3.19) List answer categories vertically instead of horizontally.

Principle 3.20) Place answer spaces consistently to either the left or right of the category labels.

Principle 3.21) Use numbers or simple answer boxes for recording of answers.

Principle 3.22) Vertical alignment of question subcomponents among consecutive questions eases the response task.

Principle 3.23) Avoid double or triple banking of answer choices.

Principle 3.24) Maintain spacing between answer choices that is consistent with measurement intent.

Principle 3.25) Maintain consistency throughout a questionnaire in the direction the scales are displayed.

Principle 3.26) Use shorter lines to prevent words from being skipped.

Step 3. Developing additional visual navigational guides, the aim of which is to interrupt established navigation behavior and redirect respondents, for example, through skip patterns.

Principle 3.27) Major Visual changes are essential for gaining compliance with skip patterns.

Principle 3.28) Words and phrases that introduce important, but easy to miss, changes in respondent expectations should be visually emphasized consistently, but sparingly.

IV. **Sampling procedures**: Finding people to answer your questions

A. Selecting randomly a sample from the accessible population

1. **Simple Random Sampling:** From one list of names, randomly choosing individuals to serve as a sample representative of the population.

2. **Systematic Random Sampling:** From one list of names, randomly choosing ***one*** individual from some fraction of the total number of individuals. The random selection of this one individual will directly determine all the remaining members of the sample. For example, if you want a sample of 10 people from a population of a hundred, you may randomly choose 1 of the first 10 people in your list. If you randomly chose person 3, the third person in every remaining group of 10 persons would be included in the study (i.e., person # 13, 23, 33, 43, 53, 63, 73, 83, and 93).

3. **Stratified Random Sampling:** From two or more list of names, randomly choosing individuals to serve as a sample representative of each population. This strategy is used when one intends to compare different groups in terms of how they responded to the survey.

4. **Cluster Random Sampling:** Not having a list of names, individuals are randomly chosen according to group membership (cluster). You may randomly choose classrooms in a school and use the students in each randomly selected classroom as your study participants. Here, we assume you have a list of the classrooms, but not a list of names.

B. Simple Random Sampling: How would I estimate the mean response to an item?

Often in survey research, we intend to calculate the average response to a Likert item or a Score determined by summing a set of related Likert items. In either case, sample means are calculated to summarize the results of the survey research. It is therefore important to gain an appreciation of how to estimate a mean.

To point out the obvious, the mean of a set of scores obtained from a sample of people is an estimate of the mean of a set of scores in the population. Sample means become more accurate in representing the population mean when larger sample sizes are used. You might think that calculating a sample mean is sufficient in estimating the population mean, but the truth is calculation of the sample mean is never enough.

The problem with a sample mean is that it is merely a point estimate. It's a point estimate because it's the best single point to estimate the population mean (the true mean). Nonetheless, on its own, it gives us no information about just how accurate it is. Said differently, it does not inform us about the degree of sampling error potentially involved in the calculation. To determine the accuracy of the sample mean, we must obtain an interval estimate. The sample mean is placed within a confidence interval so that we can say with some degree of confidence what the true mean is based on the sample mean. Usually, the confidence interval is calculated to give us 95% confidence regarding what the population mean is. Only rarely are we confident that the sample mean is the population mean, but, with a confidence interval, we can be 95% confident about what our population mean is.

Mean estimation is easiest when only one list of names is sampled from, when a simple random sample is selected.

(The following examples of Estimation obtained from Schaffer, Mendenall, & Ott's 1990 text, Elementary Survey Sampling 4th Ed.,)

1. **Mean Estimation**: An Example of the ***Simple*** Random Sample

   A federal auditor is to examine the accounts for a city hospital. The hospital records obtained from a computer show a particular accounts receivable total, and the auditor must verify this total. If there are 28,000 open accounts in the hospital, the auditor cannot afford the time to examine every patient record to obtain a total accounts receivable figure. Hence the auditor must choose some sampling scheme for obtaining a representative sample of patient records. After obtaining the patient accounts in the sample, the auditor can then estimate the accounts receivable total for the entire hospital. If the computer figure lies within a specified distance of the auditor's estimate, the computer figure is accepted as valid. Otherwise, more hospital records must be examined for possible discrepancies.

   Suppose that of N = 1,000 hospital records, the auditor draws a simple random sample of 200 records, in such a way that any one hospital record has an equal chance of being selected for inclusion in the sample.

   First, the auditor will estimate the mean average amount of money due for all 1,000 accounts.

The mean turns out to be \$94.00 and the sample variance is \$445.21. To estimate μ for all 1,000 accounts, we will use the previously identified equations.

$$S^2_{\bar{X}} = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{445.21}{200}\left(\frac{1000-200}{1000}\right) = 1.7808$$

The Margin of Error or “B” = $2\sqrt{S^2_{\bar{X}}} = 2\sqrt{1.7808} = \$2.67$

So, we may be 95% confident that the true mean of all accounts receivable falls between \$ 91.33 and \$ 96.67.

Note: When using estimation procedures the **Margin of Error** is sometimes referred to as the **Bound** on the error of estimation. This is why “B” is sometimes used to represent the Margin of Error, as in the example above.

2. **Total Estimation**: An Example of the ***Simple*** Random Sample

Using the information in the previous example, we may estimate the total amount of money due to the hospital based on the records chosen and compare this figure to what the hospital's computer says that it is.

Recall,

$$S^{2}_{\bar{x}} = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{445.21}{200}\left(\frac{1000-200}{1000}\right) = 1.7808$$

So, $S^{2}_{\hat{t}} = N^2 * \left(\frac{s^2}{n}\left(\frac{N-n}{N}\right)\right) = (1000)^2 * 1.7808 = 1780800$

and therefore,

The Margin of Error or "B" = $2\left(\sqrt{S^{2}_{\hat{t}}}\right) = 2\sqrt{1780800} = \$2668.93$

Because N=1000 and the sample mean is \$94.00, the estimated total accounts receivable due is \$94,000.

So, we may be 95% confident that our total accounts receivable due fall between \$91,331.07 and \$96,668.93. If the hospital computer reports the accounts receivable to be \$92, 447, we would say that this figure is in all likelihood a trustworthy figure.

Another example

An industrial firm is concerned about the time per week spent by scientists on certain trivial tasks. The time log sheets of a simple random sample of n=50 employees show the mean average amount of time spent on these tasks is 10.31 hours with a sample variance =2.225. The company employs N=750 scientists. Estimate the total number of man-hours lost per week on trivial tasks and place a bound on the error of estimation.

Multiplying the sample mean average of 10.31 hrs by the total number of people in the population N=750, our estimation of the total number of hours lost on trivial tasks is 7732.5 hours.

$$S^{2}_{\hat{t}} = N^2 * \left(\frac{s^2}{n}\left(\frac{N-n}{N}\right)\right) = 750^2 * \left(\frac{2.25}{50}\left(\frac{750-50}{750}\right)\right) = 23{,}625$$

$$B = 2\sqrt{S^{2}_{\hat{t}}} = 2\sqrt{23{,}625} = 307.4 \text{ hours}$$

So, we may be 95% confident that the total number of hours that the scientists actually spend on trivial tasks falls between 7425.1 hrs and 8039.9 hrs.

3. **Proportion Estimation**: An Example of the ***Simple*** Random Sample

A simple random sample of n = 100 college seniors was selected to estimate the fraction of N = 300 seniors going on to graduate school.

Say, that 15% of the 100 students surveyed indicate that they are going on to graduate school. A margin of error around this estimate would be

$$S^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right) = \frac{.15(1-.15)}{100-1}\left(\frac{300-100}{300}\right) = 0.00085849$$

$$B = 2\sqrt{S^2_{\hat{p}}} = 2\sqrt{0.00085849} = 0.0586$$

So, may be 95% confident that that actual proportion of college senior who would report going on to college falls between 0.0914 (9.14%) and .2086 (20.86%). Such a wide range may suggest that we would have been better off selecting a larger simple random sample. With only 300 seniors total in the school, it may have been reasonable to survey them all. You decide.

C. Stratified Random Sampling: How would I estimate the mean response to an item for each of several groups?

Stratification is a wonderful strategy that may fruitfully be used to obtain a more precise estimate of the population mean than simple random sampling allows. This is because when we account for group differences explicitly, we reduce the amount of error in our estimate. Stratified random sampling does not have as its objective commenting on the estimate of each group. Instead, when stratification is used, estimation of one population value becomes more sharply refined because we take into consideration group differences. Simple random sampling may nonetheless be used afterwards to color the results further.

The procedure used to estimate a population mean and build a confidence interval around the estimate actually changes when mean responses of several groups of people to an item (or complete questionnaire) are considered at once. To give you a sense of how matters change, I will provide you an example of stratified random sampling for mean responses.

Other uses of stratified random sampling will not be presented (for totals and proportions).

1. **Mean Estimation**: An Example of the ***Stratified*** Random Sample

   Consider first, the equations and contrast them with Simple Random Sampling procedure. The equations below assume three groups are being compared:

$$\overline{X}_{ST} = \frac{1}{N}\left[N_1\overline{X}_1 + N_2\overline{X}_2 + N_3\overline{X}_3\right]$$

$$S^2_{\overline{X}\ \mathrm{st}} = \frac{1}{N^2}\sum_{i=1}^{3} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$$

$$\mathrm{B} = 2\sqrt{\mathrm{S}^2_{\overline{\mathrm{X}}_{st}}}$$

$$\text{For } \overline{X}_1,\ \mathrm{B}_1 = 2\sqrt{\left(\frac{N_1 - n_1}{N_1}\right)\left(\frac{s_1^2}{n_1}\right)}$$

$$\text{For } \overline{X}_2,\ \mathrm{B}_2 = 2\sqrt{\left(\frac{N_2 - n_2}{N_2}\right)\left(\frac{s_2^2}{n_2}\right)}$$

$$\text{For } \overline{X}_3,\ \mathrm{B}_3 = 2\sqrt{\left(\frac{N_3 - n_3}{N_3}\right)\left(\frac{s_3^2}{n_3}\right)}$$

Example

A corporation that markets textbooks wishes to obtain information regarding a Western Civilization text written for 11th grade students. Teachers from three states to which the book will be marketed are randomly selected to evaluate the text for a stipend. The corporation wants to know not only how favorable the teachers rate the book using an evaluation sheet, but also want to compare teachers across the three states of interest. The score that teachers give will on a 100 point scale, with 90 and above suggesting an "A" for the text, 80 to 89, a "B", etc. The accompanying table indicates the mean ratings of the teachers sampled in each state, along with the standard deviation. Moreover, the sample sizes (n) and population sizes (N) are noted. Using the information provided, calculate the 95% confidence interval around the estimates, and indicate how favorable the teachers rated the text, and note whether any true differences are likely to exist in the opinions of teachers across the three states.

Georgia	**New York**	**California**	**Overall**
Mean = 90	Mean = 85	Mean = 70	
s = 10	s = 15	s = 12	
N = 500	N = 1000	N = 800	N = 2300
n = 50	n = 100	n = 80	

$$\overline{X}_{ST} = \frac{1}{2300}\left[500(90)+1000(85)+800(70)\right] = 80.87 \text{ or } 81$$

$$S^2_{\overline{X}\text{ st}} = \frac{1}{(2300)^2}\left[500^2\left(\frac{500-50}{500}\right)\left(\frac{100}{50}\right)+1000^2\left(\frac{1000-100}{1000}\right)\left(\frac{225}{100}\right)+800^2\left(\frac{800-80}{800}\right)\left(\frac{144}{80}\right)\right]$$

$$B = 2\sqrt{S^2_{\overline{X}_{st}}} = 2\ \sqrt{0.66385} = 1.63$$

So, we are 95% confident that the population of teachers in the three states would evaluate the text to be 79.24 and 82.50, either a low B or high C. In response, the corporation will require the authors to make revisions.

Now. we will calculate the simple random sample bound around each sample mean on the next page.

For $\overline{x}_1$, $B_1= 2\sqrt{\left(\frac{N_1-n_1}{N_1}\right)\left(\frac{s_1^2}{n_1}\right)} = 2\sqrt{\left(\frac{500-50}{500}\right)\left(\frac{100}{50}\right)} = 2.68\rangle$ 90 ± 2.68 or (87.32, 92.68)

For $\overline{x}_2$, $B_2= 2\sqrt{\left(\frac{N_2-n_2}{N_2}\right)\left(\frac{s_2^2}{n_2}\right)} = 2\sqrt{\left(\frac{1000-100}{1000}\right)\left(\frac{225}{100}\right)} = 2.84\rangle$ 85 ± 2.84 or (82.16, 87.84)

For $\overline{x}_3$, $B_3= 2\sqrt{\left(\frac{N_3-n_3}{N_3}\right)\left(\frac{s_3^2}{n_3}\right)} = 2\sqrt{\left(\frac{800-80}{800}\right)\left(\frac{144}{80}\right)} = 2.55\rangle$ 70 ± 2.55 or (67.45, 72.55)

So, we are 95% confident that Georgia teachers rate the text as either an A or B; New York teachers rates the book as a B; and California teachers rates the book as either an C or D. The teachers from the three states are different from one another. You may be thinking that there is some chance that New York teachers may overlap with Georgia teachers, given that both may give the text a B, based on our estimates. Remember that the lower bound for Georgia teachers is (90 - 2.68 = 87.32) which overlaps the upper bound for New York teachers (85 + 2.84 = 87.84). What does this suggest? It suggests that we may not be confident that the population of Georgia teachers and the population of New York teachers are truly different, with respect to their assessment of the textbook. Overlapping confidence intervals around two estimates suggests no statistically significant difference between the two group means. **Insight: Estimation can be used for testing nondirectional hypotheses!**

D. So, how large should my sample size be in a study?

Let us return to our first examples to determine how the sample size needed is identified prior to conducting a study. In each case, we need two pieces of information to plug into the sample size equation: The Population variance (or an estimate thereof) and a clear idea of how precise we want our estimate to be.

"Absurd!" you may say! To determine sample size I must pull ***out of thin air*** two pieces of information – my best guess of what the (1) **population variance** is AND (2) my decision about how precise I want my estimate to be, how small I want my margin of error (**B**) to be.

"How can I do that BEFORE I've collected any data?" Well, its actually not so hard to play around with scenarios to guess how precise you want your estimate to be. You want to be able to set the standard of how accurate your result will be. Let's say you know on your questionnaire the minimum and the maximum score will be. On a 5 item scale, consisting of Likert items ranging from 1 to 4, the largest score would be 20 (i.e., the value "4" as an option is selected by the respondent for each of 5 items); the smallest score would be 5 (i.e., the value "1" as an option is selected by the respondent for each of 5 items). Ask yourself, "If on my scale, the mean response is 10… How certain would I like to be?" Maybe you could tolerate the true score for the population being between 9 and 11, but you wouldn't want to be less certain. In this case, your margin of error (i.e., bound) is $\pm$ 1. That was easy. In fact, it was empowering because I get to decide how precise I want my estimate to be (before I even know what the mean response is going to be).

The more tricky issue is: How decide what the population variance is going to be. This is a tad stickier to explain. The short answer is that I can use a previous estimate obtained with this same population, the last time this scale was administer to a sample from the population in question. But, you say, "I just created my scale for the first time. How can I possibly know that?" Well, if you just developed your questionnaire for the first time, this approach will not do. Either you have to operate blind the first time you give your questionnaire, or make use of a delightful mathematical finding that a man by the name of Tchebysheff' cam up with. Don't let his name scare you off.

In the absence of any information about the population, how do I determine the **population variance**? Well, consulting Tchebysheff's theorem, we know that the range is often approximately equal to four standard deviations (**4σ**); so, one-fourth the range may serve as an approximate value of **σ**. We will define the **range** as (**highest value – lowest value**).

Consider two scenarios

If our Likert item has five options from which to choose, the widest possible range for the item equals 4 (i.e., 5 - 1). So, our estimate of the population standard deviation is 4/4 = 1.00. Squaring this value gives the **population variance** of 1.00.

If our total score on a scale has 100 points (and its possible for someone to get as low as a 0 score), the widest possible range for that scale is 100 (i.e., 100 - 0), and our estimate of the population standard deviation is 100/4 = 25. Squaring this value gives the **population variance** of 625.

Can you determine that the population variance could be with the 5 item scale mentioned above. Recall that each item had four response choices- The Likert item went from 1 to 4, probably Strongly Disagree to Strongly Agree.

Isn't Tchebysheff' one of the Coolest, Smokin' fellows you ever knew?

Once you have this information, you are ready to use the following formulae for determining sample size in the case of simple random sampling.

Overview for Simple Random Sampling

N = Finite Population Size	n = Sample size
B = Margin of Error	$S^2_{\text{statistic}}$ = Squared std. err. of *

Parameter	Selection of Sample	Estimation

Mean (μ)	n = $\frac{N\sigma^2}{(N-1)D+\sigma^2}$
	D = $B^2/4$
		$\overline{X}$
		$S^2_{\overline{X}} = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$
		B = $2\sqrt{S^2_{\overline{X}}}$

Total (τ)	n = $\frac{N\sigma^2}{(N-1)D+\sigma^2}$
	D = $B^2/4N^2$
		$\hat{\tau} = N\overline{X}$
		$S^2_{\hat{t}} = N^2 * \left(\frac{s^2}{n}\left(\frac{N-n}{N}\right)\right)$
		B = $2\sqrt{S^2_{\hat{t}}}$

Proportion (p)
such .50 or.80
	n = $\frac{N\sigma^2}{(N-1)D+\sigma^2}$
	D = $B^2/4$ and
	$\sigma^2 = p(1-p)$
		$\hat{p}$
		$S^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right)$
		B = $2\sqrt{S^2_{\hat{p}}}$

As promised, we'll return a previously discussed mean estimation example. To estimate sample size, I will need three ingredients: **1.** The population N, **2.** the Bound of my choosing, and **3.** a population variance.

1. **Mean Estimation**: An Example of the ***Simple*** Random Sample

A federal auditor is to examine the accounts for a city hospital. The hospital records obtained from a computer show a particular accounts receivable total, and the auditor must verify this total. If there are 28,000 open accounts in the hospital, the auditor cannot afford the time to examine every patient record to obtain a total accounts receivable figure. Hence the auditor must choose some sampling scheme for obtaining a representative sample of patient records.

How big should the sample size be?

Well, at least we know the population size. Since we're estimating money owed to the hospital, and the hospital is a pretty big institution, having 95% confidence that the estimate fall somewhere between ± $10.00 should not be a problem. The question is: How big is the population variance? Now we could take the easy way out and use what the computer says. Our only other option would be to use Tchebysheff's theorem. Tchebysheff's theorem could only be used in this case if we knew at the very least the most money owed the hospital by a single person. We know the least would be $0.00. If no one could owe more than $1000.00, our range is cautiously assumed to be from $1000.00 – $0.00 or $1000.00. The range divided by 4 is $250.00, and so our population variance is conservatively estimated to be $62,500.00 (the square of $250.00)

Using the information above and the relevant equation, we proceed as follows:

$$D = B^2/4 = (\$10.00^2)/4 = \$25.00$$

$$n = \frac{N\sigma^2}{(N-1)D+\sigma^2} = \frac{(28{,}000)62{,}500}{(28{,}000-1)25+62{,}500} = 2295.16,$$

rounded up to 2,296 patient records needed for the sample.

2. **Mean Estimation**: Another Example of the ***Simple*** Random Sample

At James Madison University, I served as a member of the SACS committee, assigned particularly to assist the university with the collection and reporting of university-wide assessment results. One element of preparing for the external accreditation committee involved conducting survey research with various JMU groups (the faculty, classified staff, etc.) to ask questions pertaining to perceptions about the university. One group of people for whom we were fashioning a unique survey was classified staff. The Director of Institutional Research showed me the questionnaire, and asked me how many members of the classified staff should he give the questionnaire? What would be a representative sample? It turned out that their were **831 members** of the classified staff. A quick review of the survey reveals that the Likert items were comprised of four options. Moreover, there was no interest in adding the item response up for a total score; instead the average response for each item was to be reported.

We decided to use a **margin of error of .20**. Furthermore, because each Likert item has four options, we divided the 4/4 for an estimated **population standard deviation of 1.0**. Plugging all information into the equation:

$D = B^2/4 = (.20)^2/4 = .01$, and so our sample size is determined to be

$$n = \frac{N\sigma^2}{(N-1)D+\sigma^2} = \frac{831\,(1)^2}{(831-1).01+(1)^2} = 89.26 \text{ or } 90 \text{ members of the Classified Staff}$$

3. **Mean Estimation**: An Example of the ***Stratified*** Random Sample

For another survey at James Madison University, we wanted to e-mail a questionnaire to the faculty in the five colleges, stratifying by college. The finite population sizes for faculty in each of the five colleges were:

1)	CISAT	111
2)	Business	242
3)	Arts and Languages	98
4)	Education and Psychology	90
5)	Science and Math	93
		634

Here's a table to consult when determining sample size for stratified samples.

Overview for Stratified Random Sampling
(Assuming _3_ groups)

N = Total Finite Population Size — N_i = Finite Population Size for each group
n = Total Sample size — n_i = Sample size for each group
B = Margin of Error — V(*) = Variance of *
w_i = a weight for each group (the sum of all weights must equal 1.00)

Parameter	Selection of Sample	Estimation
Mean (μ_{st})	$n = \frac{\sum_{i=1}^{3} N_i^2 \sigma_i^2 / W_i}{N^2 D + \sum_{i=1}^{3} N_i \sigma_i^2}$ $D = B^2/4$ $w_1 = N_1 \sigma_1 / \Sigma(N_i \sigma_i)$ $w_2 = N_2 \sigma_2 / \Sigma(N_i \sigma_i)$ $w_3 = N_3 \sigma_3 / \Sigma(N_i \sigma_i)$ $n_1 = n\,(w_1)$ $n_2 = n\,(w_2)$ $n_3 = n\,(w_3)$	$\overline{X}_{ST} = \frac{1}{N}\left[N_1\overline{X}_1 + N_2\overline{X}_2 + N_3\overline{X}_3\right]$ $S^2_{\overline{X}_{st}} = \frac{1}{N^2}\sum_{i=1}^{3} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$ $B = 2\sqrt{S^2_{\overline{X}_{st}}}$ For $\overline{X}_1$, $B_1 = 2\sqrt{\left(\frac{N_1 - n_1}{N_1}\right)\left(\frac{s_1^2}{n_1}\right)}$ For $\overline{X}_2$, $B_2 = 2\sqrt{\left(\frac{N_2 - n_2}{N_2}\right)\left(\frac{s_2^2}{n_2}\right)}$ For $\overline{X}_3$, $B_3 = 2\sqrt{\left(\frac{N_3 - n_3}{N_3}\right)\left(\frac{s_3^2}{n_3}\right)}$
Total (τ_{st})	$n = \frac{\sum_{i=1}^{3} N_i^2 \sigma_i^2 / W_i}{N^2 D + \sum_{i=1}^{3} N_i \sigma_i^2}$ $D = B^2/4N^2$ $w_1 = N_1 \sigma_1 / \Sigma(N_i \sigma_i)$ $w_2 = N_2 \sigma_2 / \Sigma(N_i \sigma_i)$ $w_3 = N_3 \sigma_3 / \Sigma(N_i \sigma_i)$ $n_1 = n\,(w_1)$ $n_2 = n\,(w_2)$ $n_3 = n\,(w_3)$	$\hat{\tau}_{st} = N_1\overline{X}_1 + N_2\overline{X}_2 + N_3\overline{X}_3$ $S^2\hat{t}_{st} = \sum_{i=1}^{3} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$ $B = 2\sqrt{S^2\,\hat{t}_{st}}$ For $\overline{X}_1$, $B_1 = 2\sqrt{N_1^2\left(\frac{N_1 - n_1}{N_1}\right)\left(\frac{s_1^2}{n_1}\right)}$ For $\overline{X}_2$, $B_2 = 2\sqrt{N_2^2\left(\frac{N_2 - n_2}{N_2}\right)\left(\frac{s_2^2}{n_2}\right)}$ For $\overline{X}_3$, $B_3 = 2\sqrt{N_3^2\left(\frac{N_3 - n_3}{N_3}\right)\left(\frac{s_3^2}{n_3}\right)}$

In our stratified JMU example, let’s assume that B = .20 (the same value we used in our simple random sample scenario). This means that D equals .01.

Calculating the weights necessary, using the Neyman allocation method, a method that considers unequally sized sub-populations, gives the following values.

$$w_1 = \left(\frac{N_1\sigma_1}{\sum_{i=1}^{5} N_i\,\sigma_i}\right) = \left(\frac{111\,(1)}{111(1)+242(1)+98(1)+90(1)+93(1)}\right) = \left(\frac{111\,(1)}{634}\right) = 0.17507$$

$$w_2 = \left(\frac{N_2\sigma_2}{\sum_{i=1}^{5} N_i\,\sigma_i}\right) = \left(\frac{242\,(1)}{111(1)+242(1)+98(1)+90(1)+93(1)}\right) = \left(\frac{242\,(1)}{634}\right) = 0.38170$$

$$w_3 = \left(\frac{N_3\sigma_3}{\sum_{i=1}^{5} N_i\,\sigma_i}\right) = \left(\frac{98\,(1)}{111(1)+242(1)+98(1)+90(1)+93(1)}\right) = \left(\frac{98\,(1)}{634}\right) = 0.15457$$

$$w_4 = \left(\frac{N_4\sigma_4}{\sum_{i=1}^{5} N_i\,\sigma_i}\right) = \left(\frac{90\,(1)}{111(1)+242(1)+98(1)+90(1)+93(1)}\right) = \left(\frac{90\,(1)}{634}\right) = 0.14195$$

$$w_5 = \left(\frac{N_5\sigma_5}{\sum_{i=1}^{5} N_i\,\sigma_i}\right) = \left(\frac{93\,(1)}{111(1)+242(1)+98(1)+90(1)+93(1)}\right) = \left(\frac{93\,(1)}{634}\right) = 0.14668$$

$$n = \frac{\sum_{i=1}^{5} N_i^2 \sigma_i^2 / W_i}{N^2 D + \sum_{i=1}^{5} N_i \sigma_i^2} = \frac{\frac{111^2(1)^2}{111(1)/634} + \frac{242^2(1)^2}{242(1)/634} + \frac{98^2(1)^2}{98(1)/634} + \frac{90^2(1)^2}{90(1)/634} + \frac{93^2(1)^2}{93(1)/634}}{634^2(.01) + 111(1)^2 + 242(1)^2 + 98(1)^2 + 90(1)^2 + 93(1)^2} =$$

$$n = \frac{70374 + 153428 + 62132 + 57060 + 58962}{4019.56 + 634} = \frac{401956}{4653.56} = 86.376 \text{ or } 87 \text{ faculty members}$$

$$n_1 = n\,w_1 = n\left(\frac{N_1 \sigma_1}{\sum_{i=1}^{5} N_i \sigma_i}\right) = 87\left(\frac{111(1)}{111(1) + 242(1) + 98(1) + 90(1) + 93(1)}\right) = 87\left(\frac{111(1)}{634}\right) =$$

15.23 or 16 faculty members from CISAT College

$$n_2 = n\,w_2 = n\left(\frac{N_1 \sigma_1}{\sum_{i=1}^{5} N_i \sigma_i}\right) = 87\left(\frac{111(1)}{111(1) + 242(1) + 98(1) + 90(1) + 93(1)}\right) = 87\left(\frac{111(1)}{634}\right) =$$

33.21 or 34 faculty members from the College of Business

$$n_3 = n\,w_3 = n\left(\frac{N_1 \sigma_1}{\sum_{i=1}^{5} N_i \sigma_i}\right) = 87\left(\frac{111(1)}{111(1) + 242(1) + 98(1) + 90(1) + 93(1)}\right) = 87\left(\frac{111(1)}{634}\right) =$$

13.45 or 14 faculty members from College of Arts and Languages

$$n_4 = n\,w_4 = n\left(\frac{N_1 \sigma_1}{\sum_{i=1}^{5} N_i \sigma_i}\right) = 87\left(\frac{111(1)}{111(1) + 242(1) + 98(1) + 90(1) + 93(1)}\right) = 87\left(\frac{111(1)}{634}\right) =$$

12.35 or 13 faculty members from College of Education and Psychology

$$n_5 = n\,w_5 = n\left(\frac{N_1 \sigma_1}{\sum_{i=1}^{5} N_i \sigma_i}\right) = 87\left(\frac{111(1)}{111(1) + 242(1) + 98(1) + 90(1) + 93(1)}\right) = 87\left(\frac{111(1)}{634}\right) =$$

12.76 or 13 faculty members from College of Science and Math

V. Survey Implementation

A. **Tailored Design Method:** A set of procedures for conducting successful self-administered surveys that produce both high quality information and high response rates.

The Tailored Design Method may also be defined as the development of survey procedures that create respondent trust and perceptions of increased rewards and reduced costs for being a respondent, that consider features of the survey situation, and that have as their goal the overall reduction of survey error.

The most important concept underlying Tailored Design has to do with applying social exchange ideas to understand why respondents do or do not respond to questionnaires. Rather than relying on one basic procedure for all survey situations, it builds effective social exchange through knowledge of the population to be surveyed, respondent burden, and sponsorship

B. **Key Terms**

1. Information organization: the prescribed order in which we want people to process words and symbols used to convey the questions and all needed instructions to respondents

2. Navigational guides: the graphical symbols and layout used to visually direct people along a prescribed navigational path for completing the questionnaire.

3. Social Exchange: A theory of human behavior used to explain the development and continuation of human interaction. The theory asserts the actions of individuals are motivated by the return these actions are expected to bring, and in fact usually do bring, from others. Three elements are critical for predicting a particular action: ***rewards, costs, and trust***.

4. Rewards: what one expects to gain from a particular activity

5. Costs: what one gives up or spends to obtain the rewards

6. Trust: the expectation that in the long run the rewards of doing something will outweigh the costs.

C. Ways of providing Rewards in light of Social Exchange theory

1. Show positive regard
2. Say thank you
3. Ask for advice
4. Support group values
5. Give tangible rewards (even token rewards like pens)
6. Make the questionnaire interesting
7. Give social validation
8. Inform respondents that opportunities to respond are scarce

D. Ways of reducing Social costs in light of Social Exchange theory

1. Avoid subordinating language
2. Avoid embarrassment
3. Avoid inconvenience
4. Make questionnaires short and easy
5. Keep requests similar to other requests to which a person has already responded

E. Ways of establishing Trust in light of Social Exchange theory

1. Provide a token of appreciation in advance
2. Sponsorship by legitimate authority
3. Make the task appear important
4. Invoke other exchange elements

F. Five Needed Elements for Achieving High Response Rates

1. Respondent Friendly Questionnaire
   a. Particularly affects item nonresponse rates more so than overall response rates

2. Four Contacts by First Class Mail, with an Additional "Special" Contact (The Implementation System Discussed in next section of Notes)

   a. First Contact: Prenotice Letter
   b. Second Contact: The Questionnaire Mailout
   c. Third Contact: The Postcard Thank You/Reminder
   d. Fourth Contact: The First Replacement Questionnaire
   e. Fifth Contact: The Invoking of Special Procedures

3. Return Envelopes with Real First-Class Stamps

4. Personalization of Correspondence

5. Token Prepaid Financial Incentives ($1.00 to $5.00)

G. Detailed Features of the Implementation System (More on the Second Element)

1. First Contact: Prenotice Letter

a. Sent to respondents a few days prior to questionnaire.

b. Indicates that a response would be appreciated.

c. Should be brief, personalized, positively worded, and aimed at building anticipation rather than providing details or conditions for participation.

d. If a small token of appreciation is to be provided with the questionnaire, it should be mentioned but without going into details.

e. It should be sent first-class mail and time to arrive only days to a week ahead of actual questionnaire.

f. Provide letterhead stationary, personalized address and signature

g. Use a letter instead of a postcard because it takes 20 seconds to get an event into long term memory.

h. If the survey is to be sent on behalf of a sponsor **or** to persons belonging to an organization or affiliated with some group, it would be useful to have the letter be sent by that sponsor, organization, or group, thereby invoking exchange elements of authority and legitimacy.

2. Second Contact: The Questionnaire Mailout

a. Cover letter: Limit to one page, written for a person with an educational level a little less than the anticipated, average survey recipient.

b. Date: Include a specific date

c. Inside name and address: Whenever possible, always include this personal information.

d. Salutations. Only use names when the gender of the recipient is known.

e. What is this letter about?

f. Why this request is useful and important. Avoid being so specific that your personal bias is detectable by the recipients. A general purpose statement is all that is needed.

g. Explain that answers are confidential, while being honest but brief. Long explanations inhibit responses.

h. Note that Participation is Voluntary, but add the request that those declining participation should "Please let us know by returning the uncompleted questionnaire."

i. Enclosures of stamped return envelope and token of appreciation may be mentioned but briefly.

j. Who to contact with questions? Provide this information preferably with a toll free number. This information conveys trust, and is an essential component of a good cover letter.

k. A real signature in contrasting ink. Use a pressed blue ball-point pen signature, preferably signed on a soft surface

l. Addition of a postscript. Consider that postscripts are highly visible to a reader, so take advantage of this fact for whatever you choose to use this (to express thanks again, include other important information).

m. Identification of each questionnaire. Numerically identify every survey unless the survey addresses a sensitive topic. If the survey questionnaire addresses a sensitive topic, include in the envelope with the survey a self-addressed post card, which confirms that the recipient has responded.

n. Inclusion of token financial incentive

1. Should the incentive be sent with the questionnaire or as payment afterwards? With the questionnaire.

2. How large a cash incentive is needed? Don't use coins; a dollar bill is recommended. Larger amounts considered in previous research have been shown to quickly level off with little to no advantage.

3. Should token financial incentives be sent as cash or as a check? Checks work about as well as cash for amounts from $5 or higher. Smaller checks may be seen as a nuisance. Recognize that many people do not cash checks, saving some money. On the other hand, processing checks costs money.

4. Will material incentives work as well? No, not nearly as well, but they have some impact.

5. Do lotteries, contributions to charities, or offers of prizes improve response rates? Lotteries, contributions to charities, or offers of prizes have much too small an affect on responses.

6. Is it worth while to repeat the incentive when a replacement questionnaire is sent? No evidence suggests this.

o. The importance of first class postage and how to apply it. Avoid bulk mailing.

p. Use a stamped return envelope.

q. Assembling and inserting the mailout package. Four components should be included: the questionnaire, cover letter, token incentive, and return envelope. All four enclosures should come out of envelope at once, and the most appealing aspect of each should be visible.

r. Selecting the mailout date. Day of the week, particular month, etc. have not shown to matter with respect to survey mailouts, although Dillman advises that the Thanksgiving to Christmas season be avoided.

3. Third Contact: The Postcard Thank You/Reminder

a. Repeated studies suggest that nearly half the return envelopes are postmarked within two or three days after being received by respondents.

b. The postcard follow-up is written to jog memories and rearrange priorities rather than to overcome resistance. The inevitably high nonresponse rate to any mailing is probably due less to conscious refusals than to either unrealized good intentions or the lack of any reaction at all.

c. It should be sent after a week, should convey a sense of importance being carefully worded not to sound impatient or unreasonable.

d. Precise wording is important. The first lines should convey in simple terms that a questionnaire was sent in the previous week and why. The second paragraph should thank those who have already returned their questionnaire and request that those who have not do so "Today". Follow this sentence up with a message indicating how important each recipient is to the success of the study as described in the original cover letter. The third and final paragraph is an invitation to call for a replacement questionnaire if one is needed. Complete the postcard with a statement of appreciation, and the researcher's name, title, and signature.

e. The respondent's name and address are individually printed out on the reverse side. The name is not repeated on the message side.

f. Send the postcard to all questionnaire recipients. This way they can be prepared in advance.

4. Fourth Contact: The First Replacement Questionnaire

a. This letter has a tone of insistence that the previous three contacts do not have.

b. Its strongest aspect is the first paragraph, in which recipients are told that their completed questionnaire has not been received.

c. This message is one of the strongest forms of personalization, communicating to respondents that they are indeed receiving individual attention. It reinforces messages contained in previous contacts that the respondent is important to the success of the survey.

d. It conveys, in a manner that is encouraging, that others have responded.

e. The social usefulness of the survey is reemphasized, implying that the usefulness of the study is dependent on the return of the questionnaire.

f. The recipient is reminded which member of the household is to complete questionnaire.

g. The letter is concluded by mentioning the enclosed replacement questionnaire, the usual note of appreciation, and now familiar blue ball-point signature.

h. It is sent by first-class mail in the same type of envelope used for the initial mailing.

i. Avoid using too strong a tone, but definitely use a stronger tone than previous contacts.

j. Consider the usefulness of adding a postscript to the letter addressing some of the feedback given by those who have completed the questionnaire.

k. Schedule the fourth contact a full two weeks or slightly longer after the postcard reminder.

5. **Fifth Contact:** Invoking Special Procedures

   a. The tone of the final contact letter should be softer than the fourth contact.

   b. The important way in which it differs from previous contacts is its packaging and manner of delivery.

   c. Certified mail has been shown in research to be substantially effective.

   d. Certified mail presents a problem: it requires a person to be home when the package is delivered. Alternatives include Priority mail and special delivery.

   e. Telephone calls: Yet another alternative to Certified Mail. Scripts are provided to interviewers, people are personally reassured about the nature of the study, thanked for their consideration and reassured that this is the last attempt at contact. Phone calls should be made within a week after the anticipated arrival of the fourth contact. Interviewers should be prepared to listen to concerns and prepared (trained) to answer questions about the survey.

H. Dynamics of the Implementation Process

1 Handling Respondent Inquiries: Each mailing is likely to bring reactions other than a completed questionnaire. It is important to be prepared to answer anticipated questions. Frequent questions, cited by Dillman, are on page 189.

2 Evaluating Early Returns: It is important to immediately open the first surveys returned to inspect whether any unanticipated problems emerge. These problems made be dealt with in latter contacts.

I. What to Do when Maximizing Response Quality Systems Seems Impossible

1. Budget constraints: What part of survey process does one cut out? It’s largely up to you. Consider what you or your stakeholders are willing to compromise.

VI. Preserving Quality: Reducing Coverage and Sampling Errors

A. Essential Definitions

1. **Survey populations** consist of all of the units (individuals, households, organizations) to which one desires to generalize survey results.

2. The **sample frame** is the list from which a sample is drawn in order to represent the survey population.

3. The **sample** consists of all units drawn from the population for inclusion in the survey.

4. The **completed sample** is all the units that return the completed questionnaire.

5. **Coverage error** results from every unit in the survey population not having a known, non-zero, chance of being included in the sample.

6. **Sampling error** is the result of collecting data from only a subset of the members in the sampling frame.

B. Reducing Coverage Error

1. Does the list (i.e., sample frame) contain everyone in the survey population?

2. Does the list include name of people who are not in the study population?

3. How is the list maintained and updated?

4. Are the sample units included on the list more than once?

5. Does the list contain other information that can be used to improve the survey?

6. What to do when no list is available? One answer: Cluster Random Sampling

C. Reducing Sampling Error: The importance of Probability Sampling (Covered previously – e.g., simple random sampling).

How large should a sample be?

VII. **Mixed Mode Surveys** Collecting information using multiple methods

A. Five Situations for Use of Mixed-Mode Surveys

1. Collection of the same data from different members of a sample using different modes. Issues here include: **(a)** whether people answer questions differently depending upon the mode used **(b)** considering how one may combine modes in the most cost-effective way and **(c)** considering how the first mode may be designed so that second mode may be more successful.

2. Collection of panel data from the same respondent at a later time. Mixing modes in panel studies is particularly challenging because one want to compare time 1 with time 2, and the change in mode may on its own affect the responses obtained. Economic efficiency often motivates researchers to employ a new mode for a panel study.

3. Collection of different data from the same respondent during a single data collection period. For example, an interview may be immediately followed by a questionnaire on which sensitive questions are asked. Also, questionnaires may be administered to identify a subset of persons who fulfill certain criteria so that they may be interviewed. These Mixed-Mode strategies actually **increase** the quality of the responses obtained.

4. Collection of comparison data from different populations. (comparing different populations using different modes). Vague quantifiers must have the same measurement properties or the results will be questioned.

5. Use one mode to prompt completion by another mode (e.g., using a telephone to prompt one to respond to questionnaire).

B. Consequences of Mixed-Mode Designs Examine Table 6.1 on page 219 in Dillan's text.

C. Why People May Answer Self-administered and Interview Questionnaires Differently

The most basic cause: People tend to design questions differently for different modes used. Unimode construction is a solution.

Why do they?

1. Social Desirability: “I care about whether you’ll judge me.”
2. Acquiescence: “I want to be agreeable, to get along.”
3. Primacy/Recency effects (Order effects)

D. Unimode Design as a Solution for Certain Mode Differences

1. Make all response options the same across modes **and** incorporate them into the survey question.

2. Avoid inadvertently changing the basic question structure across modes in ways that change the stimulus.

3. Reduce the number of response categories to achieve mode similarity

4. Use the same descriptive labels for response categories instead of depending on people’s vision to convey the nature of the scale concept.

5. If several items must be ranked, precede the ranking question with rating questions addressing each response option in the ranking question. (If you must ask a ranking question at all).

6. Develop equivalent instructions for skip patterns that are determined by answers to several widely separated items

7. Avoid question structures that unfold.

8. Reverse the order in which categories are listed in half the questionnaires. (Counterbalancing for order effects)

9. Evaluate interviewer instructions carefully for unintended response effects and consider use for other modes.

VIII. **Internet and E-mail**: Reaching out into cyberspace.

A. Internet and Interactive Voice Response Surveys

1. Surveys on the Internet

a. E-mail surveys are simpler to compose than Web surveys, but are more limited with regard to their visual stimulation and interaction capabilities, and provide fewer options for dealing with difficult structural features of questionnaires such as extensive skip patterns.

b. Web surveys have a more refined appearance, to which color may be added, but also survey capabilities far beyond those available for any other type of self-administered survey. They can be designed to provide more dynamic interaction, extensive skip patterns can be included in ways that are not visible to the respondent; pop-up instructions can be provided for individual instructions; drop-down boxes with response options can be used to provide immediate coding of answers to certain questions that are usually asked with an open-ended format; and shapes, colors and pictures may be used to improve the appearance of the survey questionnaire.

c. Current coverage is inadequate for most e-mail and web surveys.

d. Effects of computer equipment and telecommunications access

e. Effects of computer literacy

f. Computer logic versus questionnaire logic and the needed to design with both in mind.

g. A fundamental distinction between designing for paper and the Internet.

B. Design principles for E-mail Surveys

Principle 11.1: Utilize a multiple contact strategy much like that used for regular mail surveys.

Principle 11.2: Personalize all e-mail contacts so that none are part of a mass mailing that reveals either multiple recipient addresses or a listserv origin.

Principle 11.3: Keep the cover letter brief to enable respondents to get to the first question without having to scroll down the page.

Principle 11.4: Inform respondents of alternative ways to respond, such as printing and sending back their response.

Principle 11.5: Include a replacement questionnaire with the reminder message.

Principle 11.6: Limit the column width of the questionnaire with the reminder message.

Principle 11.7: Begin with an interesting but simple to answer question.

Principle 11.8: Ask respondents to place X's inside brackets to indicate their answers.

Principle 11.9: Consider limiting scale lengths and making other accommodations to the limitations of e-mail to facilitate mixed-mode comparisons when response comparisons with other modes will be made.

C. Principles for Constructing Web Surveys

Principle 11.10: Introduce the Web questionnaire with a welcome screen that is motivational, emphasizes the ease of responding, and instructs respondents about how to proceed to the next page.

Principle 11.11: Provide a PIN number for limiting access only to people in the sample.

Principle 11.12: Choose for the first question an item that is likely to be interesting to most respondents, easily answered, and fully visible on the welcome screen of the questionnaire.

Principle 11.13: Present each question in a conventional format similar to that normally used on paper self-administered questionnaires.

Principle 11.14: Restrain the use of color so that figure/ground consistency and readability are maintained, navigational flow is unimpeded, and measurement properties of questions are maintained.

Principle 11.15: Avoid differences in the visual appearance of questions that result from different screen configurations, operating systems, browsers, partial screen displays, and wrap around text.

Principle 11.16: Provide specific instructions on how to take each necessary computer action for responding to the questionnaire, and give other necessary instructions at the point where they are needed.

Principle 11.17: Use drop-down boxes sparingly, consider the mode implications, and identify each with a "click here" instruction.

Principle 11.18: Do not require respondents to provide an answer to each question before being allowed to answer any subsequent ones.

Principle 11.19: Provide skip directions in a way that encourages marking of answers and being able to click to the next applicable question.

Principle 11.20: Construct Web questionnaires so they scroll from question to question unless order effects are a concern, or when telephone and Web survey results are being combined.

Principle 11.21: When the number of answer choices exceeds the number that can be displayed in a single column on one screen, consider double-banking with an appropriate grouping device to link them together.

Principle 11.22: Use graphical symbols or words that convey a sense of where the respondent is in the completion process, but avoid those that require significant increases in computer resources.

Principle 11.23: Exercise restraint in the use of question structures that have known measurement problems on paper questionnaires, such as check all that apply and open ended questions.

Empirically Evaluating the Quality of the Survey Questionnaire

IX. **Reliability and Item analysis**: Appraising the Consistency of your Questionnaire

A. Reliability and the Classical True Score Model

1. Whenever a survey or measure is administered, the administrator of the measure would like some assurance that the survey or measure results could be replicated if the same individuals were measured again under the same circumstances.

2. **Reliability**: The consistency (or reproducibility) of a measure's scores. This consistency may be expected to occur when the same people (1) are reexamined with the same measure on different occasions, or (2) receive two different forms of the a measure on the same occasion, or (3) receive one form of a measure on the same occasion. In the latter case, you want to know how consistently examinees were in responding to all items on the measure (item homogeneity). The theory behind this is that the more consistent the examinees are in responding across items, the more consistent their performance is likely to be with future administrations.

3. The classical true score model

a. Charles Spearman was fascinated with the concept of correlation. From 1904 to 1913 he published logical and mathematical arguments that scores are fallible measures of human traits, and thus the observed correlation between fallible measure scores is lower than the correlation between their "true objective values". Out of this came the foundation for the classical true score model: $X = T + E$. The essence of Spearman's model was that any observed measure score X can be envisioned as a composite of two hypothetical components: a true score and a random error component. How many of you have taken a measure and felt that your performance on that measure truly measured your ability?

b. Definition of the true score

1) True score: the average of the observed scores obtained on an infinite number of repeated measurements with the same measure.

c. Definition of Error

1) Measurement Error:

   The discrepancy between an examinee's observed measure score and true score. In other words, E = X - T. Recall that the average of the observed scores obtained over an infinitely repeated number of measurements equals the true score. For this reason, the average error score obtained over a repeated number of measurements is expected to be zero. Put simply, given that X = T + E, whenever X = T, E must be zero. If the average of the Xs equals T, then the average of the Es will necessarily equal zero.

2) A point that needs to be made here is that there are two broad categories of measurement error: systematic and random.

3) Systematic measurement errors are those which consistently affect an individual's measure score because of some particular characteristic of the person or the measure that has nothing to do with the construct being measured. For example, on some reading measures for children, the examiner says a word and the examinee is required to circle the letter that indicates the beginning sound. A hearing impaired child may hear "bet", when the examiner says "pet" and mark an incorrect response. If the measure were repeated the child would make similar errors, and this child's scores would be consistently depressed across measurement occasions. Respondents who answer items with some response set also illustrate systematic measurement errors. Take, for instance, the person who always marks "disagree" when he finds an attitude scale item ambiguous. Because such tendencies persist across repeated measurements with the same instrument and affect the examinee's score in a constant fashion, they are systematic errors of measurement.

4) Random measurement errors are those that affect a person's score because of purely chance happenings. They may affect an examinee's score in either a positive or negative direction, adding to or subtracting from the examinee's score. Sources of random errors include guessing, distractions in the measurement situation, administration errors, content sampling, scoring errors, and fluctuations in the examinee's state. Fluctuations in an examinee's behavior may be general enough to affect overall measure performance (e.g., a headache affects performance) or may be very brief and specific (e.g., misreading a question, miscopying a math problem, or forgetting momentarily an answer.) If the examinee were to repeat the same exam, the random errors that affect his or her score on the first occasion probably would not be repeated, although other random errors would undoubtedly occur. In the equation, X = T + E, E represents random measurement error and T includes systematic measurement error.

5) Both **random** and **systematic** measurement errors are a source of concern in score interpretation. Systematic measurement errors, although consistent, may cause measure scores to be inaccurate, and thus reduce their practical utility. Random measurement errors reduce both the consistency and utility of the measure scores. It would be illogical to expect measurements to be useful if we did not have some confidence that they were consistent. Thus, measure developers have a responsibility to demonstrate the reliability of score obtained from their measures. Such demonstrations require empirical studies that are usually based on a theoretical model for describing the extent to which random errors influence the scores. Note that systematic error that is bound up with the true score actually contributes to the reliability of the measure, just as random error detracts from the reliability of the measure.

4. The standard error of measurement

Reliability is a concept that permits the measure user to describe the proportion of true score variance in a group's observed measure scores. In many situations, however, the measure user is more concerned with how measurement errors affect the interpretation of individuals' scores. Although it is never possible to determine the exact amount of error in a given score, classical measure theory provides a method for describing the expected variation of each individual examinee's observed scores about the examinee's true score. Just as the total group has a standard deviation, theoretically each examinee's personal distribution of possible observed scores around the examinee's true score has a standard deviation. When these individual error standard deviations are averaged for the group, the result is called the standard error of measurement and is denoted $\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}}$

B. Procedures for Estimating Reliability

1. Two measure administrations

a. When you took the GRE, it should have been administered under controlled conditions at a particular site on a given date. Because cheating on the exam must be controlled, examinees in adjacent seats should have taken different forms of the exam covering the same content. The question is just how fair was it to give two different forms of a measure to the examinees? Did one group receive an easier exam? or a more understandable exam? One way to answer this question is to use **Alternate Form reliability.**

b. **Alternate Form reliability**: indicates how consistently examinees respond to two similar forms of a measure (different items, similar content). The two measure forms are administered one right after the other to the same group of examinees (giving a break is OK to guard against burn-out). Then, a Pearson correlation coefficient is calculated between the scores obtained for each measure form. The result is called a **"*coefficient of equivalence*"**. The measurement errors that are the primary concern of this procedure are those due to differences in content of measure forms. Of course, errors due to administration, scoring, guessing, examinee mismarkings, and other temporary fluctuations are concerns as well. The problem associated with this procedure concerns the difficulty of making sure the two forms are truly equivalent.

c. **Test Retest reliability**: indicates how consistently the same examinees respond to a measure over time. Calculate a Pearson correlation coefficient between two administrations of the same measure and call the result a **"*coefficient of stability*"**. The measurement errors that are the primary concern of this procedure are those due to temporary changes in an examinee's state. Of course, errors due to administration, scoring, guessing, examinee mismarkings, and other temporary fluctuations are concerns as well. The problem with this type of reliability is that exposure to the measure contents promotes better performances on later administrations of the same measure (i.e., "practice makes perfect"). Moreover, if the measure administrations are separated in time long enough so that the examinees forget the measure contents, a new problem arises: maturation and outside learning most likely will occur and thus influence future measure performance. A critical question in the design of a test-retest study is this: How much time should elapse between measurements? There is no single answer. The time should be long enough to allow effects of memory or practice to fade but not so long as to allow maturational or historical changes to occur in examinee's true scores. The purpose for which the measure scores are to be used should be taken into account in designating the waiting time.

d. **Test-Retest with alternative forms**: In this case, you administer one form of the measure, wait for some specified period, then administer the other form of the measure. Such reliability coefficients tend to be smaller in value than other reliability coefficients. The correlation coefficient measuring the relationship between the two forms is referred to as a ***"coefficient of stability and equivalence"***.

2. One measure administration

a. Split-Half reliability: Indicates how much of a relationship exists between two halves of a measure. You can split a measure into two halves in one of four ways: (1) You may randomly assign items to two groups; (2) you may assign all even numbered items to one group and odd numbered items to the other group; (3) you may rank order items according to difficulty levels based on the responses of examinees and then assign odd and even numbered items to two groups; or (4) you may match items according to content and then assign items similar content to different groups. After splitting the items into two groups, you calculate a correlation coefficient between scores for the two halves. The resulting value is called a "coefficient of equivalence". The problem with this "coefficient of equivalence" is that reliability will be underestimated, smaller than it should be. You **must** correct for the underestimation of reliability caused by splitting the measure into two forms. To do so, plug the coefficient of equivalence into the Spearman

$$\rho_{XX} = \frac{2\,\rho_{AB}}{1+\rho_{AB}}$$

Brown formula). where ρ_{AB} is the correlation between the two halves. The problem with this method is that different estimates are possible depending upon the way you split the measure. Another problem is that this technique requires the two halves to be equivalent in difficulty.

How the Split half method and Spearman Brown prophesy formula are calculated.

Summary of Three Steps:

1) **Divide the measure into two equal (or near equal) parts using any one of several methods. For our purposes, we will use the odd/even method.**
2) **Calculate a Pearson product correlation coefficient between the two halves.**
3) **Plug the Pearson product correlation coefficient into the Spearman Brown prophesy formula**

Step 1) Divide the measure into two equal parts using the odd/even method.

Responses of 10 Examinees to 6 Items, Dichotomously Scored

A.	**Items**							
B.	**ODD**				**EVEN**			
Examinee	1	3	5	**Total 1**	2	4	6	**Total 2**
1	0	0	0	0	0	0	0	0
2	0	0	1	1	0	0	0	0
3	1	1	1	3	0	1	0	1
4	1	1	1	3	1	1	1	3
5	1	1	1	3	1	1	1	3
6	0	1	0	1	0	0	0	0
7	0	1	1	2	0	1	0	1
8	0	0	0	0	0	1	0	1
9	1	1	1	3	0	1	0	1
10	0	0	0	0	1	1	1	3

Step 2) Calculate a Pearson product correlation coefficient between the two halves.

Calculating two standard deviations.

Examinee	Total 1	$(X_i - \mu_i)$	$(X_j - \mu_j)^2$	Total 2	$(X_j - \mu_j)$	$(X_j - \mu_j)^2$
1	0	0 – 1.6 = –1.6	2.56	0	0 – 1.3 = –1.3	1.69
2	1	1 – 1.6 = –.6	.36	0	0 – 1.3 = –1.3	1.69
3	3	3 – 1.6 = 1.4	1.96	1	1 – 1.3 = –.3	.09
4	3	3 – 1.6 = 1.4	1.96	3	3 – 1.3 = 1.7	2.89
5	3	3 – 1.6 = 1.4	1.96	3	3 – 1.3 = 1.7	2.89
6	1	1 – 1.6 = –.6	.36	0	0 – 1.3 = –1.3	1.69
7	2	2 – 1.6 = .4	.24	1	1 – 1.3 = –.3	.09
8	0	0 – 1.6 = –1.6	2.56	1	1 – 1.3 = –.3	.09
9	3	3 – 1.6 = 1.4	1.96	1	1 – 1.3 = –.3	.09
10	0	0 – 1.6 = –1.6	2.56	3	3 – 1.3 = 1.7	2.89
Mean	**1.6**			**1.3**		
		SS =	16.48		SS =	14.1
		σ^2 =	1.648		σ^2 =	1.41
		σ =	**1.28**		**E. σ =**	**1.19**

Calculating the covariance.

Examinee	Total 1	$(X_i - \mu_i)$	Total 2	$(X_j - \mu_j)$	$(X_i - \mu_i)(X_j - \mu_j)$
1	0	0 – 1.6 = –1.6	0	0 – 1.3 = –1.3	2.08
2	1	1 – 1.6 = –.6	0	0 – 1.3 = –1.3	0.78
3	3	3 – 1.6 = 1.4	1	1 – 1.3 = –.3	– 0.42
4	3	3 – 1.6 = 1.4	3	3 – 1.3 = 1.7	2.38
5	3	3 – 1.6 = 1.4	3	3 – 1.3 = 1.7	2.38
6	1	1 – 1.6 = –.6	0	0 – 1.3 = –1.3	0.78
7	2	2 – 1.6 = .4	1	1 – 1.3 = –.3	– 0.12
8	0	0 – 1.6 = –1.6	1	1 – 1.3 = –.3	0.48
9	3	3 – 1.6 = 1.4	1	1 – 1.3 = –.3	– 0.42
10	0	0 – 1.6 = –1.6	3	3 – 1.3 = 1.7	– 2.72
Mean	1.6		1.3		Sum = 5.20
					Covariance = **.52**

$$\rho\ x_i x_j = \frac{covariance}{\sigma_i\ \sigma_j} = \frac{.52}{1.28(1.19)} = .34$$

Step 3) Plug ρ_{AB} (.34) into the Spearman Brown prophecy formula

$$\rho_{xx'} = \frac{2(\rho x_i x_j)}{1+\rho x_i x_j} = \frac{2(.34)}{1+.34} = .51$$

b. <u>Kuder-Richardson reliability procedures </u>(i.e., KR 20 or KR21): Both the KR 20 or KR21 procedures determine how internally consistent the items are and do not require a split between measure halves. ***The Split half procedure has the disadvantage of giving different estimates of reliability depending on which way you split the measure in half!***

The KR 20 or KR21 procedures, conversely, improve upon the Split half procedure by essentially splitting the measure into as many pieces as there are items. Both procedures accomplish this by producing what amounts to being a summary index of all the correlations that exist among the items. Both procedures are appropriate when used for dichotomously scored items (i.e., when you have right-wrong answers). The names of the two procedures were taken from the numbered steps in the derivation in the journal article within which they were presented.

$$KR_{20} = \frac{k}{k-1}\left(1 - \frac{\Sigma pq}{\hat{\sigma}^2_x}\right) \qquad KR_{21} = \frac{k}{k-1}\left(1 - \frac{\hat{\mu}(k - \hat{\mu})}{k\,\hat{\sigma}^2_x}\right)$$

For both the KR_{20} and the KR_{21} "k" represents the number of items in the measure, σ^2_x,the total measure variance. For the KR_{20}, pq is the variance for and item scored 1 or 0. Note that the summation indicates that the variance of each item must be computed and then these variances must be summed for all items. The KR_{21} is different than the KR_{20} in that it assumes that all items are of equal difficulty. Both procedures are usually used to assess the reliability of test scores and are seldom appropriate for analyzing questionnaire data.

How the KR20 and KR21 are calculated.

Summary of Four Steps:

1) **Calculate the item variances, and add them up.**
2) **Calculate the variance for the measure scores (i.e., the items summed for each person).**
3) **Note how many items you have on the measure. (6 items)**
4) **Plug the appropriate value into the KR20 or KR21 formula.**

Responses of 10 Examinees to 6 Items, Dichotomously Scored

C.	**Items**						
Examinee	1	2	3	4	5	6	Total[a]
1	0	0	0	0	0	0	0
2	0	0	0	0	1	0	1
3	1	0	1	1	1	0	4
4	1	1	1	1	1	1	6
5	1	1	1	1	1	1	6
6	0	0	1	0	0	0	1
7	0	0	1	1	1	0	3
8	0	0	0	1	0	0	1
9	1	0	1	1	1	0	4
10	0	1	0	1	0	1	3
p_j (Item Mean)	.40	.30	.60	.70	.60	.30	
q_j or $(1 - p_j)$	.60	.70	.40	.30	.40	.70	
p_jq_j (Item Variance)	**.24**	**.21**	**.24**	**.21**	**.24**	**.21**	

Step 1) Calculate the item variances, and add them up.

$$\Sigma \boldsymbol{pq} = .24 + .21 + .24 + .21 + .24 + .21 = \mathbf{1.35}$$

Step 2) Calculate the variance for the measure scores (i.e., the items summed for each person).

	Total Score (X)	(X − μ)	(X − μ)^2
1	0	0 − 2.9 = − 2.9	8.41
2	1	1 − 2.9 = − 1.9	3.61
3	4	4 − 2.9 = 1.1	1.21
4	6	6 − 2.9 = 3.1	9.61
5	6	6 − 2.9 = 3.1	9.61
6	1	1 − 2.9 = − 1.9	3.61
7	3	3 − 2.9 = .10	.01
8	1	1 − 2.9 = − 1.9	3.61
9	4	4 − 2.9 = 1.1	1.21
10	3	3 − 2.9 = .10	.01
	Sum = 29	$SS_X = \Sigma(X-\mu)^2 =$	40.9
	μ = 2.90	**$\sigma^2 = {}^{SS}/_{N} = {}^{40.9}/_{10} =$**	**4.09**

Step 3) Note how many items you have on the measure. (ANS: 6 items)

Step 4) Plug the appropriate value into the KR20 or KR21 formula.

k = the number of items = 6

Σpq = sum of the item variances = 1.35

$\hat{\sigma}^2_x$ = the variance of the total measure scores= 4.09

$\hat{\mu}$ = the mean of the total measure scores = 2.90

$$KR_{20} = \frac{k}{k-1}\left(1-\frac{\Sigma pq}{\sigma^2}\right) \qquad KR_{20} = \frac{6}{6-1}\left(1-\frac{1.35}{4.09}\right) = 0.80391198044$$

OR …

$$KR_{21} = \frac{k}{k-1}\left(1-\frac{\mu(k-\mu)}{k(\sigma^2)}\right) \qquad KR_{21} = \frac{6}{6-1}\left(1-\frac{2.90(6-2.90)}{6(4.09)}\right) = 0.760391198044$$

c. Cronbach's Alpha (i.e., Coefficient Alpha): this procedure also determines how internally consistent the items are and does not require a split between measure halves. This procedure is different from the Kuder-Richardson reliability procedures because it is not restricted to right-wrong responses, but may be used for polytomous responses as well. The Cronbach's alpha represents the most general case for internal consistency, and so it may be used instead of the Kuder-Richardson or the coefficient of equivalence corrected by the Spearman-Brown.

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i{}^2}{\sigma_X{}^2}\right)$$

How Cronbach's Coefficient Alpha is calculated.

Summary of Four Steps:

1) Calculate the item variances, and add them up.
2) Calculate the variance for the measure scores (i.e., the items summed for each person).
3) Note how many items you have on the measure. (6 items)
4) Plug the appropriate value into the Coefficient Alpha formula.

Responses of 10 Examinees to 6 Items, Polytomously Scored

	Items						
Examinee	1	2	3	4	5	6	Total[a]
1	1	1	2	3	2	1	10
2	4	5	4	5	4	5	27
3	2	4	5	4	5	3	23
4	3	4	4	3	4	3	21
5	3	3	3	3	3	3	18
6	1	2	2	1	1	4	11
7	2	2	3	2	1	2	12
8	3	3	3	4	3	3	19
9	3	4	4	3	3	4	21
10	2	1	2	3	2	2	12
Item Variance	0.84	1.69	0.96	1.09	1.56	1.20	30.64

Step 1) Calculate the item variances, and add them up. Notice that Cronbach's coefficient alpha includes item variances that are not Σpq. This is because Cronbach's coefficient alpha may accommodate items that are NOT dichotomously scored, but are instead polytomously scored. It may, however, accommodate dichotomous items as well.

$$\Sigma\sigma^2_i = .84 + 1.69 + .96 + 1.09 + 1.56 + 1.20 = \mathbf{7.34}$$

Step 2) Calculate the variance for the measure scores (i.e., the items summed for each person).

	Total Score (X)	(X – μ)		$(X - \mu)^2$
1	10	10 – 17.4 = –7.4		54.76
2	27	27 – 17.4 = 9.6		92.16
3	23	23 – 17.4 = 5.6		31.36
4	21	21 – 17.4 = 3.6		12.96
5	18	18 – 17.4 = 0.6		0.36
6	11	11 – 17.4 = –6.4		40.96
7	12	12 – 17.4 = –5.4		29.16
8	19	19 – 17.4 = 1.6		2.56
9	21	21 – 17.4 = 3.6		12.96
10	12	12 – 17.4 = –5.4		29.16
	μ = 17.4		$SS_X = \Sigma(X - \mu)^2 =$	306.4
			$\sigma^2 = {}^{SS}/_{N} = {}^{306.4}/_{10} =$	**30.64**

Step 3) Note how many items you have on the measure. (ANS: 6 items)

Step 4) Plug the appropriate value into the Coefficient Alpha formula.

k = the number of items = 6

$\Sigma\,\sigma^2_i$ = sum of the item variances = 7.34

σ^2_x = the variance of the total measure scores= 30.64

$\hat{\mu}$ = the mean of the total measure scores = 17.4

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_X^2}\right) \qquad \alpha = \frac{6}{6-1}\left(1 - \frac{7.34}{30.64}\right) = 0.9125326370758$$

Notice that Cronbach's coefficient alpha resembles the KR_{20}. This is because both procedures accomplish the same end, though Cronbach's coefficient alpha is a more general case than the KR_{20} in that it includes the sum of ordinary population variances, NOT Σpq. You see, the variance of polytomously scored items must be calculated the ordinary way and NOT by using Σpq.

3. Interrater reliability: For some types of instruments only one set of items is used (a list of behaviors on a behavioral checklist), but multiple observations are collected for each examinee by having two or more raters complete the instrument. In this case, the consistency of the observations over raters may be of interest. Generalizability is the best procedure in this circumstance.

4. Factors that affect reliability coefficients

   a. **Group homogeneity**: The degree to which examinee scores obtained on a measure are alike. It is apparent that the magnitude of a reliability coefficient depends on variation among individuals on both their true scores and error scores. Thus, the homogeneity of the examinee group is an important consideration in measure development an measure selection. Because a measure will be more homogeneous for some groups than other groups, it stands to reason that reliability estimates will vary according to whom you give the measure.

      1) This suggests that a measure is never "reliable" or "unreliable". Contrary to common belief, it cannot be said that a measure has a reliability of .85, for example. Reliability is rather a property of the scores on a measure for a particular group of examinees. Thus, potential measure users need to determine whether reliability estimates reported in measure manuals are based on samples similar in composition and variability to the group for whom the measure will be used.

   b. **Time Limits**: When a measure has a rigid time limit such that some examinees finish, but others do not, an examinee's working rate will systematically influence his/her performance. Thus, variance in the rates at which people work becomes a part of the true score variance.

      1) Speeded measures: On some measures, the measure constructor's goal may be to assess the ability to perform the tasks rapidly.
      2) Power measures: On other kinds of measures, time limits should be long enough to allow all, or nearly all, examinees to finish

   c. **Measure Length**: The reliability of a measure is more likely to increase as number of items increase.

5. Using error estimates in score interpretation

   a. Reliability is a concept that permits the measure user to describe the proportion of true score variance in a group's observed measure scores. In many situations, however, the measure user is more concerned with how measurement errors affect the interpretation of individuals' scores. Although it is never possible to determine the exact amount of error in a given score, classical measure theory provides a method for describing the expected variation of each individual examinee's observed scores about the examinee's true score. Just as the total group has a standard deviation, theoretically each examinee's personal distribution of possible observed scores around the examinee's true score has a standard deviation. When these individual error standard deviations are averaged for the group, the result is called the standard error of measurement and is denoted $\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}}$

   b. **What's the True Score most likely?** When a standard error of **measurement** is multiplied by a critical value (say, 1.65 or 1.96), we obtain a confidence interval that suggests how confident we can be that a person's true score lies within some interval. If a person's true score is 70 and the measures standard error of measurement is 5 and we use 1.95 for a 95% confident interval then we can be roughly 95% confident that the student's ***true score*** lies somewhere between 60 and 80.

   c. **Within what range is the next Observed score obtained most likely to fall?** A concept related to the standard error of measurement is the standard error of the **estimate** $\sigma_E = \sigma_X \sqrt{1 - \rho^2_{XX'}}$. The standard error of the estimate is more useful than the standard error of measurement when talking with students, parents, and clients because you can discuss the score that the examinee could be expected to achieve if a particular ***alternative form*** of the measure were taken.

      When a ***standard error of the estimate*** is multiplied by some critical value (say, 1.65 or 1.96), we obtain a confidence interval. This confidence interval suggests how confident we can be that if a person is re-measured on a parallel measure form, her score on the second measure will lie within some interval. Suppose that a person's true score on some measure is 70, and that the measure's standard error of the estimate is 6.6. Assuming a 1.95 critical value is used (for a 95% confident interval), we can be almost 95% confident that if this person were re-measured on a parallel measure form, her ***observed score*** would lie somewhere between 66.8 and 83.2.

C. Item Analysis

Item Discrimination: The purpose of many surveys and tests is to provide information about individual differences on the construct purportedly measured by a survey or test. In either case, the parameter of interest in selection of items must be an index of how effectively the item discriminates between examinees who are relatively high on the criterion of interest and those who are relatively low. At times there is no more adequate measure of the construct available than the total survey or test score itself. In this circumstance, then total score on all the items is used as an operational definition of the examinee's relative standing on the construct of interest. The item-total (corrected) correlations are used as indicators of how well a given item relates to the construct in question overall. Standard statistical software produce these along with the reliability procedure results.

D. Introduction to Generalizability Theory (NOT ON TEST)

1. Although reliability procedures based upon classical true score theory are powerful, they are not flexible enough to accommodate all reliability problems that arise in mental measurement.

2. **Consider the following Three cases:**

    a. **Consider** when you have **more** than **two** measurements (e.g., > two sets of measure scores or ratings). Consider the possibility of a test-retest-retest procedure or having three alternative measures. What if you have three raters who together assess some performance.

    b. **Consider** when you may **not** assume that your measurements are parallel, a condition necessary for classical true score theory's reliability coefficients and standard error of measurements. Otherwise, if you do **not** meet the assumption of parallel forms, the reliability coefficient and standard error of measurement calculated are likely to be inaccurate. (Parallel measures possess equal true scores and error variances in the population while meeting five basic assumptions: (1) $X = T+E$; (2) $E(X) = T$; (3) $\rho_{ET} = .00$, (4) $\rho_{E1E2} = .00$, and (5) $\rho_{E1T2} = .00$).

    c. **Consider** when you have **more** than **one** source of error variation (i.e., more than one facet) and want to differentiate or break into chunks the total error component. (Item variation is typically the source of error in classical true score theory)