Test Procedures for this Course
Mickie Swisher, January 2020

We will use simple techniques in this course because this is not a course in data analysis (statistical or qualitative) and I do not want us to spend much time learning how to conduct and interpret the results of factor analysis or coding or any other specific procedure. I would encourage you to learn about other more elaborate techniques to use in your own research. I only techniques discussed in any detail are those that are not well described in other readings used for this class. I developed this cheat sheet because I could not find a simple explanation of the technique that I thought would be useful to you. All of these techniques are well described in the methodological literature and I provide additional resources about every one of them at the course website. Note that you can use many of these techniques at various points in the development and testing of an instrument described in the document Approach to Measurement. I will dictate a set of steps for the group and partner project. You will need to develop your own protocol for developing and testing the instrument you choose to make in the independent project.

Expert Review

This test is one of the most important things you can do to make sure that the ***content*** of your instrument is full and complete – no matter what kind of instrument you are using (scale or interview, for example). Conduct this test with people who are experts in one of two senses. (1) They are experts in the content (topic, theory) of the research. (2) They are experts in research methodology. The best alternative is to conduct an expert review with at least one person with each type of expertise.

Expert review is your first defense against collecting the wrong information and is the best way of ensuring face validity. There are no step-by-step instructions for this. You have to determine what you need. None of the other techniques to assessing validity, either quantitative or qualitative in nature, provide you with the same critical insight for so little effort on your part. Factor analysis of test data, for example, will reveal whether an instrument that is supposed to measure three different dimensions of a construct actually does produce three distinct measurements. However, factor analysis does not tell you whether they are the *right* three measurements. Prepare a document that you will use to guide an individual through a systematic review process. Focus on the methodological issues. Do **not** go "item by item" asking for corrections to wording, grammar and such. These problems will arise naturally during both expert review and cognitive testing. Here are examples of what you may want to know.

- Does the person think your definitions of dimensions in the construct differentiate clearly between them?
- Do the dimensions identified for each construct "make sense"?
- Do the items seem to capture the concept?
- Have you failed to include some dimension or failed to include a wide enough array of items?
- Are your items "balanced" in a way that will allow you to capture the full variance of ideas in the population?
- Is your process for creating a composite score adequate?

Cognitive Test

Cognitive testing is critical to producing valid, reliable results. Your overall goal is to ensure that your instructions or protocols and the items in an index convey what you want to the people who will complete your instrument. Therefore, you conduct this test with members of the target population or people who are very much like the target population in terms of traits that could affect how they interpret and respond to your questions – like language. Consult the Collins, the Willis, and the Morse

readings for ideas about how to conduct a good cognitive test. There are several other resources available at the course website as well. Do **not** ask respondents to answer the questions in your instrument. You are not collecting data at this point. You are looking for systemic problems in your instrument. There are two general objectives for cognitive testing – improving instrument structure and ensuring that respondent and researcher share meanings.

(1) You want to make sure that the order or structure of the instrument reflects the thinking process of respondents. This is especially true for data collection methods like interviews or focus groups, but also applies to other "check the box" type instruments. We usually create interview or focus group protocols based on an order of topics that is logical for us – the researchers. This may be very different from the order in which other people think through a set of related topics. I struggle with this continually. Cognitive testing is one of a few procedures that will really help you figure out these structural problems with your instruments. It is critical because people's ability to respond effectively is constrained if the structure is poor, which can have a major impact on the measurement validity of your information. Willis gives some excellent procedures to use (think aloud, for example) that involve "talking through" the instrument with the testers.

(2) You want to know whether your interpretation of the question(s) and respondents' interpretations are equivalent. The objective of cognitive testing is to understand how people interpret the questions **(*what they think you are asking*)** and how they process information to arrive at an answer **(*the mental steps they use to create a response*).** You want people to "imagine how they would answer" questions. You may therefore want to ask people to peruse the questions and stop at sections where they experience confusion or conflicts as they think about how they would answer. Cognitive testing requires probing on your part to understand what people think you are asking them. You are really trying to understand the mental processes (cognition) that they experience in trying to answer your questions..

Whatever you do, do ***NOT*** go through every single item in a rote way asking reviewers to assess each item or asking the same question each time, like "Did that make sense to you? Do you have any changes in wording you would suggest?" Cognitive testing is time consuming for you and hard work for the reviewers. Use the time to find substantive flaws. See Willis (2005) for specific examples of how to conduct a cognitive test. I recommend an iterative procedure. Start with ONE test. I usually find that I get a lot of information in that very first test. Revise your instrument. Now test with someone else. Revise again if needed. I keep doing this until I either get very little information. In many cases, only a few cognitive tests are sufficient. However, take this advice under caution. I am rethinking my own approach to this based on some recent research about the effects of sample size on the information secured through cognitive testing (see Blair & Conrad, 2011). Several readings for the week we discuss operationalizing constructs discuss how to conduct cognitive testing (Collins, 2003; Castillo-Diaz & Padilla, 2013; Priede & Farrall, 2011).

Use a variety of techniques in the test to garner input about your instrument. Too often, people simply ask someone to "see if you can complete this questionnaire." I recommend that you ***always*** use at least four procedures in a cognitive review. (1) Explain ahead of time that this is a ***test*** of the instrument and make sure the testers understand that you are ***NOT*** collecting data – just testing the instrument. Be explicit. Explain that it is a test of your instrument – not their knowledge or experience. (2) Watch people as they review the instrument. Body language, expressions, and hesitation can help you identify problematic areas in your protocol or instrument. (3) Probe and focus on parts of the instrument that you suspect will be difficult for people to answer. (4) Focus on whether they ***interpreted*** your questions the same way you do. For example, I might have a focus group question like "Do you think the economy is improving?" This sounds straightforward, but it is not. One person might have a family member who just lost his/her job off and be thinking about that. Another person might be thinking about

the Dow Jones Index and stock performance. In short, make sure that your question leads people to think about what you want them to consider.

Cognitive testing will also reveal weaknesses in the structure of the instrument. Respondents can tell you, for example, if there are abrupt changes in topic that cause cognitive dissonance. The test can also reveal systematic problems in wording, like the use of a phrase that has an implied meaning that is disruptive to providing reliable, valid information. One year a group in this course was trying to measure "coping strategies" used by graduate students to deal with the many demands on their time, the conflicts that arise, etc. Unfortunately, they made a fundamental assumption – that all graduate students experience extreme, prolonged stress. All of their questions used phrases that implied they wanted information about conditions of high stress. They should have asked about coping strategies related to different levels of stress, treated as dimensions of stress with a variable score for each. Use cognitive testing to identify these kinds of problems.

Cronbach's Alpha, Item-Total Correlation, and Inter-Item Correlation

Principle. Cronbach's alpha is a measure of the degree to which people's answers to a set of items that are supposed to measure the same thing are coherent, that is reflect a similar position, or opinion, or feeling, or knowledge of the topic.[1] Cronbach's alpha can be used only with multi-item measures for ***the same construct or dimension of a construct*** and there must be a theoretical justification for assuming that responses to the set of items will be related and therefore similar (Gliem & Gliem 2003, p. 87). It is also limited to "check the box" instruments – especially scales, indices and tests. Note that the response format has to exhibit "equal scaling." That is the units of measurement and the "distance" between scores must be the same for all items. Simply put, this means use the same response format, such as a five-category scalar response format, for all items and make sure that they all include the full range of response (like from really, really negative to really, really positive). The "names" of the categories could differ, but all of the items have to provide five choices.

Procedure. This is a very simplistic explanation of "how it works" and I do not claim that I understand how these calculations are performed in SPSS or any other statistical package. Assume you have a sample of respondents who answered a set of 20 items that you think are related and therefore can be used as a multi-item measure in your study. However, you know that responding to 20 items is onerous for respondents and you suspect that a much smaller set of items would provide just as much information as 20 items. You want to eliminate the "not so good" items. One way to do this is to calculate the split-half correlation.[2] To calculate a split-half correlation, you divide the **items** into two halves -- in this case two groups of 10 items each. You calculate the correlation coefficient between the responses from everyone in your sample to Half 1 and Half 2 of the items. If all the items really are "pretty much measuring the same thing," the correlation coefficient should be high. Cronbach's alpha goes further to combine many different split-half correlations. Cronbach's alpha is the **average of the correlation coefficients when the 20 items are divided into every possible combination of two sets of ten**. That is many correlations, not feasible to do by hand, but easily done by computer and available on most analytic software. It gives precise results and is a common procedure. Interpreting

[1] Cronbach's alpha can be considered a measure of reliability or of validity. Both usages are in the literature. In my view, the alpha statistic is primarily a measure of reliability (consistency), but the accompanying scores like the item-total correlation are, in my opinion, closely related to validity. At any rate, reliability and validity are inherently related. An unreliable measure is by definition cannot yield a valid result.

[2] This is ***not*** the same procedure as the dual sample or split-sample correlation discussed under test-retest reliability below. The dual or split **sample** correlation is based on dividing the **respondents** into two halves. Cronbach's divides the ***items*** into halves**.**

the results of Cronbach's alpha is straightforward, but it does require some decision-making. I will walk through one complete example. This is tedious, but may be helpful.

Table 1 provides the output from an analysis. Look first at the Cronbach alpha statistic at the top of Table 1 (Chronbach alpha: 0.76). Like test-retest, there are no firmly established or even agreed upon standards for what Cronbach alpha should be. You have to decide. As in the case of test-retest scores, most people would say that anything above 0.90 indicates excellent consistency among items, anything above 0.80 as good consistency, and most would consider anything above 0.70 acceptable. However, researchers accept a Cronbach's alpha as low as 0.60 and you will have to make this decision.

Table 1. Output for Cronbach's alpha and associated statistics – all items included

Summary for Scale Mean = 59.48 Std. Dv. = 9.23 Valid N: 40 Cronbach alpha: 0.75 Standardized alpha: 0.76 Average inter-item correlation: 0.20					
Variable	**Mean if Deleted**	**Variance if Deleted**	**Std. Dv. if Deleted**	**Item-Total Correlation**	**Alpha if Deleted**
Var. 1	54.71	75.23	8.67	0.46	0.73
Var. 2	55.46	72.64	8.52	0.47	0.72
Var. 3	54.78	74.23	8.62	0.50	0.72
Var. 4	55.06	69.54	8.34	0.61	0.71
Var. 5	55.26	70.75	8.41	0.55	0.72
Var. 6	55.49	70.84	8.42	0.57	0.71
Var. 7	55.02	72.37	8.51	0.55	0.72
Var. 8	55.23	71.17	8.43	0.53	0.72
Var. 9	55.16	71.12	8.43	0.54	0.72
Var. 10	55.19	62.96	8.54	0.46	0.73
Var. 11	55.45	81.66	9.04	0.05	0.77
Var. 12	56.54	79.45	8.91	0.18	0.75
Var. 13	55.21	80.94	8.99	0.08	0.76
Var. 14	57.28	83.07	9.11	0.01	0.77
Var. 15	56.87	84.67	9.20	***-0.07***	0.78

If the overall Cronbach's alpha were your only concern, you would make the decision based purely on this initial statistic. However, even if the overall Cronbach alpha is acceptable to you, you still need to look at how the individual items (called variables in most statistical packages) affect the Cronbach's alpha. Look at the column labelled "Alpha if Deleted." This tells you what the Cronbach's alpha score would be if you delete each item. For example, if I delete Var. 1 (item 1), Cronbach's alpha will ***decrease*** to 0.73 (from 0.75). If I delete Var. 11, it would ***increase*** to 0.77. In this example, deleting variables 11, 12, 13, 14 and 15 would all increase Cronbach's alpha.

However, there are two more factors to consider in conjunction with the change in alpha. The most important with regard to consistency is the **item-total correlation** because it is a better measure of the cohesiveness of the items than Cronbach alpha alone. Look at the column labelled "Item-Total Correlation." This is the correlation between any one item (Var. 1 in Table 1) and the summed score of all other items (Vars. 2-15 in Table 1). In Table 1, the item-total correlation for Var. 1 is 0.46. Many researchers feel that an item-total correlation of more than 0.50 is acceptable, and perhaps even 0.40. Again, there are no firm rules. However, with 15 items remaining, you should try to improve both the Cronbach's alpha and the item-total correlation. Look for items with negative or very low item-total correlations. These low or negative correlations are telling you that people responded differently to these items than they did to the remainder of the items. Think of these low/negative numbers as

indicating that the person responded “out of character” to the item – typically meaning that the item did not provide a consistent measurement of the construct of interest.

Based on Table 1, you would definitely eliminate Var. 15 (negative item-total correlation), Vars. 11, 13 and 14 are weak, and item 12 may be problematic. However, do not eliminate several items all at once. Remember that the item-total correlation is the total between on single item and the summative score for all other items. Every time you remove an item, you change the summative score. E.g., everything changes when you remove one item – including Cronbach’s alpha and every item-total correlation. The best practice is to remove the item with the lowest correlation coefficient (Var. 15 in our example) and then rerun the test. Then if there is still an item with a low item-total correlation and whose removal would increase Cronbach alpha, remove that item.

The last important number is the **inter-item correlation**, 0.20 in our example (first row in Table 1). This is ***average correlation*** of all participants’ responses to one item and every other item, e.g. the average correlation in responses to Var. 1 and Var. 2, Var. 1 and Var. 3, etc. This number is a sort of “average” correlation among responses to individual items. A low inter-item correlation is of concern, but not as much as a low item-total correlation. In general, improving the item-total correlation should also improve the inter-item correlation.

I deleted Vars. 14 and 15 in the second iteration (run) of the data, then Vars. 11 and 13 in the third iteration. The results of the third iteration (Table 2) shows how removing some items changes all of these values. Cronbach alpha is 0.87 and the inter-item correlation increased from 0.20 to 0.40. However, look at the Item-total correlation column in Table 2. In this iteration, the item-total correlation for Var. 12 has become ***negative*** and this is the only item whose deletion would increase Cronbach alpha. That may seem logical because after it was “questionable” in the first iteration. However, it was not certain that the item-total correlation for Var. 12 would decrease when I eliminated other items. I have seen just the opposite occur. In short, this procedure is quick and easy in almost every statistical package. Complete separate iterations to remove items or at most remove a couple of items with negative/very low correlation coefficients at each iteration.

Table 2. Output for Cronbach’s alpha and associated statistics – after iteration three

Summary for Scale Mean = 46.37 Std. Dv. = 9.23 Valid N: 40 Cronbach alpha: 0.87 Standardized alpha: 0.87 Average inter-item correlation: 0.40					
Variable	**Mean if Deleted**	**Variance if Deleted**	**Std. Dv. if Deleted**	**Item-Total Correlation**	**Alpha if Deleted**
Var. 1	41.60	73.78	8.59	0.53	0.86
Var. 2	42.35	71.88	8.48	0.50	0.86
Var. 3	41.68	73.05	;8.55	0.55	0.86
Var. 4	41.95	66.68	8.17	0.74	0.85
Var. 5	42.16	68.78	8.29	0.63	0.86
Var. 6	42.38	68.25	8.26	0.69	0.85
Var. 7	41.91	70.25	8.38	0.65	0.85
Var. 8	42.12	67.64	8.22	0.70	0.85
Var. 9	42.06	67.70	8.23	0.70	0.85
Var. 10	42.08	68.44	8.27	0.67	0.85
Var. 12	43.44	83.67	9.15	***-0.02***	0.90

Table 3 shows the final iteration (iteration 4) after deleting Var. 12. Cronbach's alpha has increased to 0.90, which is excellent. The average inter-item correlation has increased to 0.48, which is acceptable, and the lowest item-total correlation is for Var. 2 at 0.50. Further, Cronbach's alpha has stabilized. In fact, it changed only very slightly between iterations two and four – e.g. eliminating Var. 14 and 15 had a very big impact on Cronbach alpha, but iterations three and four caused only slight changes. This kind of stabilization in the Cronbach alpha is a good indicator that you have identified and removed items that do not contribute to the consistency of your measure. Cronbach alpha tends to be higher with a larger number of items (just due to the nature of the calculations that produce it). Therefore, when you can delete items and leave the alpha either unchanged or increase it, this is a very positive indication that you are narrowing down the items to the essential few that you want to include in the instrument.

Table 3. Output for Cronbach's alpha and associated statistics – after iteration four

Summary for Scale Mean = 43.44 Std. Dv. = 9.17 Valid N: 40 Cronbach alpha: 0.90 Standardized alpha: 0.90 Average inter-item correlation: 0.48					
Variable	**Mean if Deleted**	**Variance if Deleted**	**Std. Dv. if Deleted**	**Item-Total Correlation**	**Alpha if Deleted**
Var. 1	38.66	72.61	8.52	0.54	0.89
Var. 2	39.41	70.98	8.43	0.50	0.90
Var. 3	38.74	72.08	8.49	0.55	0.89
Var. 4	39.02	65.48	8.09	0.76	0.88
Var. 5	39.22	67.57	8.22	0.64	0.89
Var. 6	39.45	67.26	8.20	0.69	0.88
Var. 7	38.97	69.23	8.32	0.66	0.89
Var. 8	39.18	66.58	8.16	0.70	0.88
Var. 9	39.12	66.71	8.17	0.71	0.88
Var. 10	39.15	66.97	8.18	0.70	0.88

Limitations

This procedure is limited to instruments that use at least ordinal data in the individual responses. Note that it does ***not ensure that the content of the instrument is complete or even "right."*** It tells you whether a group of items provoke consistent responses, which may mean that they "measure the same thing." You could have very good Cronbach's alpha scores, excellent item-total correlation values, etc. – and still be measuring the "wrong things." This is one weakness with all of the more "mechanical" or technical approaches to assessing validity. Unfortunately, there is a tendency to assume that a "good score" on such tests means you have a "good instrument." This is not necessarily true. You need to use an array of procedures to ensure validity. Reliability (stability and consistency) is a requirement for validity but does not ensure validity.

Tests of Discriminatory Power

We will discuss these in class. There are a number of statistical procedures that you can use. I will show you how to do one very simple test that requires little knowledge of statistics.

Other Techniques

Test-Retest

Principle. This procedure rests on the idea that a person should give the same answer to the same item on two different occasions, an assessment of *stability.* To conduct this test, you get a sample of people to respond to the items on one occasion and then the same people respond to them again on another occasion. You would expect exactly the same answer if I asked a factual kind of question (how many children do you have). However, as we have seen, most of the things we want to measure in social science are complex and involve people's feelings, ideas, or assessments about something. Therefore, some instability is normal. Do not expect 100% consistency. The question is whether a particular item gives enough consistency to warrant its use or not. You can use this procedure with all types of instruments, but is most useful with scales, indices, and other "check the box" kinds of instruments. It is not nearly as useful with methods of data collection that use narrative responses because the potential for recall influencing the retest responses is high. People spend mental time and energy developing narrative responses and therefore tend to remember what they felt and thought when retesting.

Procedure. I will focus on techniques that you can use with ordinal and interval data (not narrative, nominal, and not things counts). One simple measure is to ***determine the correlation coefficient between the responses for everyone (the entire sample) for each item***. Use Pearson's product moment correlation. The range is from 0.00 (no reliability) to 1.00 (perfect reliability). You have to determine what is acceptable and there are no rules you can use. Obviously, a coefficient of less than 0.50 indicates very poor stability. A coefficient of more than 0.90 indicates high stability. You will find some values in the literature, for example that 0.90 to 0.80 is "good," 0.80 to 0.70 "acceptable," 0.70 to 0.60 "questionable," and so on. However, these are ***not firm rules and cannot be used as absolutes.*** To some degree, it depends on what you are measuring. In some cases (efficacy of a drug), stability is critical and a coefficient of less than 0.90 is unacceptable. However, people's feelings and opinions can vary considerably even over short periods (days or weeks). We often see this in polls of public opinion where scores related to an emotional topic will change very quickly. You need to think about the degree of emotion and feeling associated with what you are measuring. The more emotionally charged the topic, the more intrinsic variance in people's responses over time you can probably (not always) expect. In general, longer periods between test and retest are probably (again not always) more subject to this natural variation. Decide what you consider acceptable and eliminate the items that show a correlation coefficient less than this cutoff score.

Limitations. The biggest concern with using test-retest is that asking the same people the same question on two occasions can generate positive response bias. This occurs when people's response on the first test causes people to respond similarly on the second test. This occurs unconsciously as well as consciously – people do not necessarily do it on purpose. Positive response bias creates an overestimate of stability. This is more problematic when the two occasions are close in time. There are three ways to reduce this threat. (1) Use many items to comprise the variable. Starting with many items is desirable for many reasons, one of them being the reduced threat of positive response bias in all testing. Simply put, the more items people respond to, the less likely they will recall or respond to the answer they gave the first time. (2) Extend the period between tests. This is often not very practical because you are trying to create an instrument and you are probably on a tight timeline. (3) Use a dual- or split-sample correlation. Get a sample of say 30 people. Divide the sample ***randomly*** into two groups. Each group takes the test. The ***random assignment*** into groups is done to make sure that any trait affecting the response is "evenly distributed" in the two groups (the basis for true experiments). Assess the correlation between the two "halves" of the sample of 30 (15 in group 1 and 15 in group 2). This requires no time interval between tests. A combination of (1) and (3) would, for most things I measure, give me a reasonable estimation of stability.

Inter-Judge or Inter-Coder Measures

Principle. This procedure is useful with methods of data collection where the researcher must either make or interpret observations or responses. The procedure is most useful for establishing the stability of measures secure when collecting data through direct observation of people's activities or methods using open response formats. A researcher has to make many decisions about how to record what s/he sees – we will see examples of this when we study direct observation, interviews and focus groups. For example, the researcher has to organize, code and interpret people's responses to interview questions, even if s/he uses computer assisted coding because the researchers has to establish the parameters for the codes. All of this involves judgment on the researcher's part. The problem is to make as sure as possible that you are producing a reliable interpretation of what the participant said or did. It is easy to confuse what you *think* a person means with what the person actually wanted to tell you in an interview. Inter-judge or inter-coder measures of reliability compare the interpretations of two or more (preferable) observers. Two or three people are unlikely to arrive at the same erroneous interpretation. By comparing what multiple observers "think happened" or "think the respondent meant" you can identify a common set of standards to reduce researcher bias or error.

Procedure. Get at least two, three would be better, four better yet, people to watch what people do or take notes during a focus group or code the ideas that emerge during an interview. Compare their interpretations. When there are discrepancies, each individual needs to explain his or her "thinking" that led to the interpretation. Use this information to develop a set of rules or indicators that all agree upon for making each required judgment. Basically, you are creating a ***rubric*** to ensure uniformity in judgments. As you know, I use a grading rubric for every assignment in this class. This is to address stability in my grading procedures, to make sure that I do not use different definitions of excellent, acceptable and poor responses on different days due to variations in my own mood or ability to focus on grading, and to make sure that I consider exactly the same attributes or factors of each person's responses. An ideal procedure is to have more than one person – preferably at least three – to make all judgments and use the most common judgment as the "true" observation. The problem is that rarely will you have the personnel to do this. It is highly recommended that you do get at least two people to assist you in the first few observations or interviews or coding sessions so that you can develop rubrics.

Limitations. No matter how detailed the rubrics you develop, it is impossible to eliminate some inconsistency in judgment, just as it is impossible to eliminate all inconsistency in how people answer a question. Another limitation is that some people are simply better observers. Training helps, but I am not sure any amount of training will eliminate these inherent differences among researchers. It is also true that some researchers have a great deal of trouble separating out what they "believe" they will hear or see and what others actually say or tell them. This ability to separate "actual" from "my bias" is one aspect of critical thinking. Practicing identifying one's own biases and assumptions can improve ability. If you consistently find that other observers or interpreters reach different conclusions than you, you may want to consider using techniques of data collection and analysis that require fewer judgments.

References

Adcock, R. & Collier, D. (2001) Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review 95(3),* 529-546.

Bhattacherjee, A. (2012) *Social Science Research: Principles, Methods and Practices*. Textbooks Collection. Book 3. Available athttp://scholarcommons.usf.edu/oa_textbooks/3

Blair, J. & Conrad, F.G. (2011) Sample size for cognitive interview pretesting. *Public Opinion Quarterly* 75(4), 636-358.

Brewer, C.J. & Jones, R.L. (2002) A five-stage process for establishing contextually valid systematic observation instruments: the case of Rugby Union. *The Sport Psychologist 16,* 138-159.

Collins, D. (2003) Pretesting survey instruments: an overview of cognitive methods. Quality of Life Research 12(3), 229-238.

Castillo-Diaz, M. & Padilla, J.L. (2013) How cognitive interviewing can provide validity evidence of the response processes to scale items. Social Indicators Research 114(3), 963-975.

Gliem, J.A. & Gliem, R.R. (2003) *Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales.* Presented at the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, The Ohio State University, Columbus, OH, Oct. 8-10, 2003. Available at http://www.ssnpstudents.com/wp/wp-content/uploads/2015/02/Gliem-Gliem.pdf

Priede, C. & Farrall, S. (2011) Comparing results from different styles of cognitive interviewing: "verbal probing" vs. "thinking aloud." International Journal of Social Research Methodology 14(4), 271-287.

Willis, G.B. (2005) *Cognitive interviewing: A tool for improving questionnaire design.* Sage, Thousand Oaks, pp. 273-298 (Appendices).