Approach to Measurement
M. E. Swisher, Revised December, 2017

This document provides an approach to measurement validity that draws heavily on the ideas of Watt and van den Berg (2002) and of Adcock & Collier (2001). The Adcock & Collier document is available on e-reserves for my research methods class and is available through the UF library system by looking up the reference. You can use a direct link to get the Watt & van den Berg document http://www.cios.org/readbook/rmcs/rmcs.htm This cheat sheet does not replace or substitute for either of those two documents, nor is it a summary of what they say. The Watt & van den Berg document is required reading for this class. Adcock & Collier provide a more thorough discussion and I incorporate some of their ideas here, but I warn you in advance that is a dense reading. This cheat sheet describes the approach to measurement we will use in this course. In many regards, it is a combination of the ideas from Watt and van den Berg, drawing on Adcock and Collier. Make sure it all makes sense to you. These are the fundamental, core concepts for this class.

Background Terminology and Information Critical to Success in this Class

Constructs

Defining the constructs in a theory is critical to successful research. Even when a term is commonly used, the theoretical definition differs from the “everyday” use of the term in most cases. You want a theoretical definition of the term, not this general definition. The definition of a construct also differs from theory to theory. Power is an example. In some theories, power means “the ability and means to influence others.” In some, it means “the authority to compel others to comply.” In yet others, power means control over others. Yet, in some theories, authority, power and control are treated as different constructs, and some theories refer to legitimate and illegitimate authority. The only way to make sure your measurements are valid is to select a specific theory for your research and then create a detailed definition of the constructs in the theory, based on the research literature. You must use one theory for all of the constructs you employ in a study unless your purpose is to compare theories in which case you need to employ constructs specific to each theory as they are defined in each theory.

Examine the literature using your theory in detail. For example, the theory of planned behavior (TPB) incorporates a construct called “attitude.” An English dictionary defines attitude as a “mental position with regard to a fact or state (e.g., a helpful attitude) or a feeling or emotion toward a fact or state.” Yet, if you look at TPB (see link from course website for the first week of class), attitude is defined in TPB as an “overall assessment of a behavior.” It has two dimensions: (1) behavioral belief, which is the belief that behavioral performance is associated with certain attributes or outcomes, and (2) evaluation, which is the value attached to a behavioral outcome or attribute. E.g. in TPB attitude does not mean a “feeling or emotion.” It really means a mental assessment of what will happen if you engage in a behavior based on whether you think doing so will result in some outcome you want and also measuring whether that outcome is very important to you or not. In other theories, attitude does mean a feeling, state of mind, mental state, or emotional state. Ultimately, you have to define the constructs. You have to be create a precise definition and I want you to follow the procedure outlined in Watt and van der Berg. Do not generally expect to find a single definition for a given construct in the literature in most cases.

Variables

Constructs and how they are defined are universal to the theory. They do not vary based on topic of a study or where the study is conducted, although there are continual discussions about the definitions and they do evolve over time as a theory is used, tested, and refined. On the contrary, variables are specific to a study. You can get ideas about the variables that others have used to represent a
construct or the dimensions of a construct from the literature, but you must consider context before you try to use what others have developed because variables are specific to a given research context and study. On the other hand, all of the different variables used to represent the same theoretical construct should measure “the same abstract idea.” Otherwise, we cannot compare results from different studies. Striking the right balance between this need to define variables specific to a context while making sure they represent a theoretical construct in the way intended is a major challenge in social scientific research. For example, assume you want to understand the relationship between social status and self-efficacy. If you conduct a study in the United States, you might have three variables representing social status – characteristics of the community where the person lives, professional achievements, and public recognition. The variable public recognition would not be useful in a study conducted in a place where public recognition is rare or not valued. However, someone doing research there could develop other indicators of social status and, if both of you took care to ensure that your variables produce valid, reliable information, you could compare your results about the relationship between social status and self-efficacy – which is the critical aspect of your contribution to the body of knowledge.

Most research calls upon only a few theoretical constructs in a given theory, not all of them. You need at least one variable to represent each construct and many would argue, as I do, that you need one variable for each dimension of a construct. Therefore, you can have several variables that represent one construct if the construct has several dimensions. To measure attitude about losing weight in the example of the theory of planned behavior, you have to develop a variable for each of the two dimensions of the construct – the person’s evaluation of the likelihood of achieving an outcome (losing weight) and how important the outcome is to the person. However, you may want to use several variables to represent a single construct or a single dimension of a construct for several reasons. As we saw in the brief discussion of validity, you can increase your confidence that the items you use to ask about a specific theory are the “right” ones by having two or more measures of the same construct in your study. In short, there is no one to one correspondence between variables and constructs or even between variables and individual dimensions of constructs. You can always combine scores in both qualitative and quantitative analyses, so it is better to err on the side of redundant variables. Ultimately, you have to make the decision about how many variables, how to define them, and what they represent and you must justify your decisions in terms of the reliability, validity and discriminatory power of the scores that result (see below).

Items or Indicators

Items are the individual questions we ask people (or sometimes statements or direct observations of people that we make). Sometimes you find items in the literature, too, but usually not. As you can imagine, creating the right items to ask about something like “power in the household” is not easy. Further, the terminology is very poorly used. Most authors never refer to indicators, but rather to variables. However, some authors refer to items as variables or refer to items as indicators or may lump all of these together and call them measures. There is no universally agreed upon terminology. Be careful in your use of these terms and please use the definitions provided here for this class.

The Terminology Mess and What We Will Use

Construct = a theoretical concept as defined by the theorists who use the theory. We will use the terms construct and theoretical construct interchangeably. You will also see concept, systematized construct, and systematic construct in the literature.

Variable = the name you give a set of items or indicators that you believe taken as a whole capture the meaning of a theoretical construct. Variables may also refer to non-theoretical information – demographic descriptors, for example.

Measurement Validity -- 2
**Items or Indicators** = the *specific questions you ask or observations you make*. There is a long discussion in the literature about direct and indirect indicators and Bryman touches on this (p. 164). He gives the example of income. As a direct measure, it simply means earnings from all sources. As an indirect indicator, this is one of several items in a multi-item variable, wealth. We will use the term item and avoid the discussion about direct and indirect indicators.

**Measure, measurement or score** = the actual value, which can be a number, a code, or a theme, that emerges when people respond to items, you make observations, or you code the responses given in an interview. Note that *it is the number or value or score for a variable that is reliable, valid, or powerful for determining differences (or not).* E.g., these terms refer to the data that we produce, not the items or variables that we create and not the study itself. Think of the variables and items as a way to generate a measure.

**Qualitative Versus Quantitative Discussion**

We will use the same *conceptual approach* to try to create reliable, valid and discriminating measurements of both numerical (quantitative) and narrative (qualitative) information. Many contemporary research design and methodologists find the quantitative versus qualitative discussion a distraction that does little to improve research. Bryman’s book illustrates the problematic nature of this distinction with its two chapters on interviews – one in the “quantitative” methods section and one in the “qualitative” section. The difference in the two is that one uses a closed response format and the other an open response format for questions. I would argue, as do many, that the distinction is a false one for three reasons. First, any data can be transformed and presented as either quantitative or qualitative. I can turn open responses into categories and then use statistical procedures to analyze the data. I can take any set of numbers and reduce it to nominal (yes/no or presence/absence) data. Second, contemporary data analysis techniques blur the lines between qualitative and quantitative analyses. MAXQDA and NVivo, for example, are sophisticated qualitative data analysis software programs that include coding, aggregation of data, queries, and visualization (graphic presentations) of the data. We can and often do use qualitative data analyses to create qualitative models that are essentially the same as regression models. Third, and most important, to some degree or the other almost all social data are fundamentally qualitative in nature. It is our treatment of the data, a decision that the researcher makes, that creates the distinction between quantitative and qualitative data and mostly this has to do with how we analyze the data. This statement from Battacherjee (2012, 44-45) states this well:

> “Note that many variables in social science research are qualitative, even when represented in a quantitative manner. For instance, we can create a customer satisfaction indicator with five attributes: strongly dissatisfied, somewhat dissatisfied, neutral, somewhat satisfied, and strongly satisfied, and assign numbers 1 through 5 respectively for these five attributes, so that we can use sophisticated statistical tools for quantitative data analysis. However, note that the numbers are only labels associated respondents’ personal evaluation of their own satisfaction, and the underlying variable (satisfaction) is still qualitative even though we represented it in a quantitative manner.”

**In this course, talking about the differences between qualitative and quantitative data is NOT useful. It will often result in reduced points (scores) on assignments.**

**Single Vs. Multi-Item Variables**

In single-item variables, one item = one variable = one score. Single item variables are common in descriptive research, particularly research that lacks a theoretical basis. For example, the kinds of Measurement Validity -- 3
surveys or polls conducted prior to an election use single item variables – like “which of the following four candidates do you plan to vote for?” Questionnaires also often use single item variables (age, gender, race), particularly for descriptive statistics about the sample or population of interest. We will not spend any time on these very simple variables in this course because they are easy to create and rather standardized for many things like household income categories.

We will concentrate on **multi-item variables** in this class for three main reasons. The most important is that it is virtually impossible to measure theoretical constructs with a single item. That is like asking someone, on a scale of 1 to 10, what is your socio-economic status? It will not work. A second reason is that multi-item variables give you a composite (summative) score or measurement for each variable, which makes it possible to conduct more sophisticated qualitative and quantitative analyses of your data than single-item scores. A composite score means that you take all the answers to individual items included in a specific variable (like control over finances) and combine them through mathematical operations to create a score for the variable. The third reason is that multi-item variables allow you to use various procedures to assess things like the consistency of responses for individuals in the sample. This, in turn, helps protect against threats like expectation response bias. Most of the techniques for enhancing reliability, validity and discriminatory power require multi-item variables.

**The Central Role of Theory in Social Scientific Research**

<table>
<thead>
<tr>
<th>Operationalization</th>
<th>is the process of going from an abstract theoretical concept or construct to a list of questions or items that people can answer.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Bad process = bad information = meaningless results</em></td>
<td></td>
</tr>
</tbody>
</table>

Most social scientists base their research on theory because theories are proposed explanations of various types of phenomena and social scientists conduct research in order to be able to understand and explain the social world. Therefore, theories are the foundation for social research. You must use a theory as the basis for your work in this class. Theories consist of a number of very abstract concepts or ideas (called constructs) and a set of proposed connections between them (called linkages). The entire schema constitutes the proposed explanation or theory. The challenge for the researcher is that s/he must turn an abstract idea, a theoretical construct like socio-economic status or resilience, into questions that people can answer. We cannot just walk up to someone and ask: “On a scale of 1 to 10, what’s your socioeconomic status?” and expect to get a meaningful answer. Even if we are asking open response questions as in an interview it is not reasonable to expect that an individual can describe his/her socioeconomic status to you very well. Most people do not usually think about their experiences in abstract social scientific terms. Therefore, our task is to create a set of questions (more generally called items) that people **can understand and answer.** For socioeconomic status, these items might be things like these. “What is your household annual income? What is the profession of the head of this household? What is the highest level of education either of your parents achieved? Do you own your own home?” This process is operationalization. There are three key components in operationalization, reliability, validity and discriminatory power.

**Reliability of Measures**

The most basic component in operationalization is **reliability.** Reliability has two components.

One is **stability.** Stability means that you get the same result (answer, score) when you conduct the same measurement on the same person on two or more occasions. The most common way to assess stability is test-retest, e.g. you actually administer the same set of items to the same person on two occasions.

Measurement Validity -- 4
The second component is **consistency**. Consistency for quantitative measures means that each individual responds similarly to a set of items that are supposed to measure the same thing, e.g. that individuals are consistent in their responses. We expect different individuals to give different responses to any given question, but one individual should give us the same answer or similar answers if asked two or more highly related questions. For example, assume I ask you which of five issues (like poverty, the economy, defense, social justice and the environment) is most important to you and you say “the environment.” I later ask you to rank five proposed uses for federal tax money (to increase assistance for the poor, create jobs, fund military build-up, provide funding for tutoring for socially disadvantaged school children, and fund research to combat climate change). If the items in the second list represent the same constructs as those in the first list, you should rank “fund research to combat climate change” first in response to the second question. Finally, I might have a third question that asks you to name the single most important thing you believe the federal government should do over the next decade using an open response format. You might respond something like “reduce fossil fuel use in the U.S.” or “stop new oil drilling in the U.S.” I would probably code both of these answers as “environment priority” or something like that in qualitative analysis. In this example, you have given three responses to three highly related questions that a consistent. I asked how important different goals of government are to you in three different ways, and every time you gave me the same answer. We will use techniques like Cronbachi’s alpha and inter-rater reliability to assess reliability. Common procedures for measuring consistency in quantitative measures are Cronbachi’s Alpha, which provides a numerical score for inter-item reliability. In the case of quantitative measures, consistency means that different individuals (the researcher and two assistants, for example) agree about what the information means, e.g. that these three people assign a given open response answer to an interview question to the same theme or topic. This is inter-observer or inter-rater reliability. These and other specific techniques are described in the required readings, the supplemental materials, Bryman and in my “cheat sheet” on techniques.

Reliability is important in and of itself and it is a **prerequisite for validity of measurements**. Many of the techniques that I describe in this cheat sheet focus on the intersection of reliability and validity. For example, a “high” Cronbach’s alpha indicates that people respond in a similar pattern to a set of items. That means that the set of items are a “reliable” measure, but it also provides support of validity because we would not expect people to respond to the questions very consistently if the items represented fundamentally different constructs. Consider all of the techniques that I discuss and that we see in the literature from both perspectives – evidence of reliability and evidence of validity.

**Validity of Measures**

The second component of operationalization is **validity**. The discussion of validity is confusing. Different authors use the same term to mean different things. Some authors talk about many different forms of validity like content validity, criterion validity, face validity, and construct validity, while others make no distinctions. Authors do not agree on the definitions in many cases, and some of the definitions are not very helpful. Much of this confusion arises from epistemological differences. We will use a somewhat simplistic approach that I **want you to use for purposes of operationalizing constructs in this class**. However, I also want you to be able to use the research methodology literature that discusses other ideas about validity. Therefore, I provide resources including some required readings that use other terminology and other definitions of validity in social scientific research. I grade assignments based on how well you can use the concepts and procedures I present and how well you can discuss the strengths and weaknesses of these approaches based on the many perspectives in the methodological literature. Despite the confusion, **there is general agreement that validity in science means making sure that what you measured actually captures the meaning of the theoretical concepts or constructs of interest – as you defined them – and are appropriate in the context of your research.** In short – did you measure what you actually needed to measure to answer your research question? Do the measures or scores that result accurately capture the Measurement Validity -- 5
associations between constructs in the theory you used, as you defined them? Did your questions make sense to the people answering the questions and were they appropriate to the context (topic and factors like language and cultural norms)?

**Face Validity**

**Face validity** is by far the easiest to understand. Very simply put, face validity means that the questions seem like reasonable things to ask given the topic of the research and the context (like language or culture). E.g. “On the face of it, does this item make sense in terms of whether it’s actually relevant to the constructs of interest in my study and whether people can understand and respond to it?” See the diagram on p. 18 in Watt & van den Berg.

Here is an example. Assume I want to understand how food insecurity affects people’s self-respect and self-confidence. I obviously have to ascertain what food insecurity is for people. I cannot simply ask: “On a scale of 1 to 10, how food insecure is your family?” People cannot answer that question. If I ask “do you get food stamps,” people can answer “yes” or “no,” but most experts will say that this simple measure is not a complete measure of food insecurity because people can be food insecure and not eligible for food stamps. Therefore, I might also need to ask other things: do you sometimes go without food or skip meals because you do not have enough food, do you worry that you will be without food, do you have to ask friends and family to loan you money or help pay for groceries. On the other hand, the question “do you eat out at least once a week” is probably not a good item either. It has to do with food and we certainly hear many negative comments about how poor people could save money and eat better if they would just not buy from fast food outlets. However, even very food insecure people may do so because they hold two jobs, have many other obligations, and just do not have time to cook. So eating out at the local fast food joint may have nothing to do with food security status for many people. This is a true example. There is no clear and clean definition of food insecurity. Obviously if my study is about the relationships between “food insecurity” and “self-respect,” I have to be very careful in deciding what to ask people.

You must also consider the context for a study. The meaning of food insecurity is very different in the U.S. and in other nations. Even within one country, context matters. If you want to know about income, you have to make sure your questions are relevant for the social and economic context. For example, people living in poverty probably do not have investments that yield income, do not have enough money in the bank to draw interest, and do not have jobs that give them performance bonuses. Very wealthy people, on the other hand, may have many income streams. What you ask to determine something as simple as “income” therefore requires quite a lot of thought about the context. The target population in your small group project is “college graduate students” precisely because I think you have experience with the “context of being a graduate student.” I am trying to make context easier for you.

A lot of the literature discusses face validity as separate from content validity in general. That is not very helpful in my opinion. I prefer the approach of Watt and van den Berg who treat face validity as distinct from measurement validity. There are few explicit discussions of how to ensure face validity, despite its obvious importance. One way to think about this is that the researcher must be sure that the questions they ask people capture the actual experiences of people. This is the overlap between the world of phenomena or experience and the world of words as shown in the diagram on p. 18 in Watt and van den Berg. Very simply – did you ask the right things?

Two groups of people can provide insights about the face validity of research instruments. The first are *topical* experts. You need to know whether experts in the topic of your research agree that your items capture the needed content to answer your research question(s). Go to this group first. For the food insecure population, for example, we might get an expert panel to review the topical adequacy of a Measurement Validity -- 6
set of items. The next step is harder. In this class, you will conduct expert reviews with both topical and methodological experts. This is a critical first step in instrument development, the most basic of procedures. You also have to make sure that laypersons can make sense of your items and that they feel that the items, taken as a whole, “capture what you are trying to get at.” This group consists of people who will actually have to answer your questions, members of the target population. We will address this primarily through cognitive testing techniques. This is harder than getting expert input because the members of the target population may not be used to thinking about abstract ideas like “self-confidence” or “victimization.” Nonetheless, this is critical to creating reliable, valid instruments. I chose graduate students as the target population for your small group project to make this easier.

**Measurement Validity**

Some authors use the term measurement validity, but others do not. I prefer the term because it distinguishes between the procedures we use to make sure that our instruments do capture in a general sense the experiences and ideas that are of interest to us, as defined theoretically and procedures that we use to ensure that the individual measurements themselves are valid.

**Concurrent or congruent validity** refers to the degree to which two measures (like two questions or two sets of items) of the same or related concepts produce similar scores or results. For example, self-esteem, defined as a psychological trait, and self-confidence, defined as one’s ability to perform a behavior successfully, are closely related concepts. If I have a measure (score, value) for each of these for a number of individuals, each individual’s scores for self-esteem and for self-confidence should co-vary. Statistically the two scores should have a high, positive correlation coefficient for individuals. In qualitative analysis, the ideas representing each construct that emerge in an interview should receive assigned the same or similar codes, categories or themes for any given individual. This refers to scores for a single individual — the scores for different individuals will differ. We are concerned with the **pattern of correspondence between scores, not the actual values.** Simply put, people who score low on self-esteem are likely to score low on self-confidence. One way to assess congruent validity is to include multiple different measurements of a single construct in your research. A second way is to compare the scores (results) of your proposed measurement to a well-established measure. I expect you to use these techniques in this class.

**Divergent or discriminant validity** is just the opposite of concurrent or congruent validity. Divergent validity refers to the degree to which measures of unrelated or different constructs produce different scores. Again, in simple terms measurements of different phenomena should produce different results. You can assess divergent validity in two ways – by the degree to which scores for different constructs vary in your own instruments and by the degree to which scores that you produce for a construct differ from those of related, but theoretically dissimilar constructs in the literature. Statistically the scores for the two should have a low positive correlation coefficient or a negative correlation, or the ideas representing each would end up under very distinct codes in qualitative analysis. Again, this refers to the scores for individuals. There will be variation across individuals.

**Construct validity** requires that the measurements you use produce both concurrent and discriminant validity. However, you also need to ensure that (1) you have a full and complete measurement of the construct and (2) you have not included unrelated ideas. Three procedures are critical to establishing construct validity – in addition to establishing concurrent and discriminant validity. (1) You have to define the constructs clearly and precisely and your items must capture all of the components in the definition. (2) You have to review the literature and examine the degree to which the relationships between theoretical construct and empirical data match (or not) in the existing empirical evidence. It is surprising how often this simply is not true in the sense that the predicted matching is weak or even absent in the empirical data (see Bhattacherjee, 2012). This is why it is so important to define the dimensions (or components or whatever you want to call the related “pieces” of a broad, highly abstract Measurement Validity -- 7
construct) and create a distinct variable score for each one. (3) You need to test the degree to which the individual measurements of what you believe to be related dimensions co-vary. If the measures produce similar scores, you have evidence that all of them are measuring "the same idea." On the contrary, if one or more scores demonstrates very limited co-variance, you may reach one of two conclusions. One is that it measures "something else," not the construct of interest. The other is that you somehow made a mistake in things like item wording, item order, questioning route, or other aspects of instrument development that we discuss in this course. To do this, you must measure two or more theoretically linked constructs or dimensions of constructs in one study. You can run very simple tests like Spearman’s Rank Correlation to see if an "expected theoretical relationship between constructs" appears in your test data. That is one reason why you have to operationalize two constructs in the small group project. In "real life research," you would use things like logistic regression or structural equation modelling to examine the empirical patterns.

I realize this is confusing. Some authors refer to this as theoretical validity, also sometimes called pattern matching or nomological validity, and I occasionally use these terms. They all refer to the degree to which the actual scores for variables “match” the relationships between constructs in the theory. In essence, do your data make theoretical sense? Bhattacherjee (2012) provides the basis for this idea about validity in Figure 4.1 on page 27 of her chapter on Theories in Scientific Research. She ties construct (an abstract idea) to variable (an empirical observation) and indicates that the two should mirror each other and that the variables should bear the same relationship to each other as the constructs they represent. Adcock and Collier (2001) use the term nomological validity. W. Trochim in his Research Methods Knowledge Base (www.socialresearchmethods.net) explains the idea of pattern matching. All of these approaches refer to the ensuring that the relationships between constructs that have been well established through previous research reflected are reflected (matched) by the relationships between the scores for variables in a specific study.

Summary of Key Ideas and Processes

Examine the results of previous research. That’s the easiest and best way to get a start on discriminant and convergent validity. I do ask you in all of your assignments to look at the literature and figure out how other people measured the same things you want to measure. Base your decisions about the content of your instruments on what they learned. If someone found that “self-regulation” seems to have three dimensions instead of two, you can include the three dimensions as three variables in an assignment and see if they provide a better measure of self-regulation than just two measures.

Mixed methods research is a powerful approach to improving validity, especially construct validity. For example, in the small group assignment you will create two indices to provide measures of two different constructs. In the partner project, you will develop an interview protocol that further explores one of these constructs. This is one of the great strengths of mixed methods research – you get to test for convergent and discriminant validity because you have not only multiple measures of the same constructs, but quite different kinds of measures.

Use redundant measures. Ask about the same construct in two or three (hence the term triangulation) different ways. One student of mine wanted to know how the Cherokee perceive of their traditional agriculture compared to modern agriculture. She had a set of dyadic comparisons that asked them about the degree to which various practices, crops, etc. are part of traditional Cherokee agriculture versus modern agriculture. That produced one score. She had a summative open response question like “Overall, for you, what are the practices that differ most between traditional Cherokee agriculture and modern agriculture? This was a second score based on qualitative analysis. Finally, she had two indices that asked about the degree to which various principles (like living in harmony with nature and Measurement Validity -- 8
achieving control over nature) and reasons for farming (like making a profit or staying close to the land) are important for Cherokee farmers. This gave a third score. Convergence would mean that she got “similar answers” to all of these different measurements from individuals. Note that her measures also get at “divergent validity” to some degree. We would not expect that a person who scores high on principles associated with traditional agriculture (living in harmony with nature) would then say that applying inputs like pesticides and fertilizer are strong components of traditional Cherokee agriculture.

Having said all this, I do not want you to focus on different “kinds” of validity for two reasons. First, it is a very convoluted discussion and tends to confuse more than clarify. Adcock & Collier (2001) found 37 different terms attached to the term validity. Second, however many “kinds” of validity one might argue exist and however they might differ from each other, the goal is to get a final measurement (quantitative or qualitative) that is a “real and useful measure” of a theoretical construct applied to your topic in your context. I want to see evidence that you understand the importance of the process in your assignments and that you can apply the ideas discussed here and in the various readings to evaluate your own instruments.

**Discriminatory Power**

The third component of measurement is **discriminatory power**. This refers to the degree to which measurements can accurately capture the differences among respondents well enough for you to distinguish between respondents. Most social scientific researchers want to distinguish between individuals or groups. We want to be able to use the results of a study to assign people to categories (high versus low resilience, different stages in the developmental process, identity). Rarely discussed in the methodological literature, this is one of the most demanding aspects of creating research instruments. Discriminatory power obviously rests on adequate reliability and validity, but also requires capturing the **full range of possible responses or score on a variable**. A surprising number of research instruments fail to meet this requirement. For example, we often evaluate training events based on self-reported change in knowledge using an ordinal response format. We may ask “How much did your knowledge improve as a result of this training session?” and ask people to respond on a scale of 1 to 5 where 1 means no change and 5 means very significant improvement. The problem with this measurement is that some people may actually know less after a training session than before if the material is presented poorly or is misunderstood by the participant. Although rarely discussed, discriminatory power is a very important attribute of scores. All of the statistical tests of central tendency like ANOVA are based on the idea of within group versus between group variance. So are most qualitative forms of data analysis. We will use tests of item-total correlation to identify which of a set of items that show good face validity are most valuable for distinguishing between respondents.

**General Procedures or Approach for this Course**

The remainder of this document provides some general procedures that we will use in this course to improve reliability, validity and discriminatory power. Use this document in conjunction with required readings for the class, the cheat sheet on techniques linked at the course website, and the additional resources provided each week.

**Define all constructs, including those that are not the focus of your research, fully before you attempt to create an instrument**

A clear, detailed definition of each construct in the theory that you use for research is critical to success in research. Define all of the constructs. Make sure you clearly identify all of the dimensions for each construct. You need to define all of them in order to make sure that there is no overlap or ambiguity in the definitions. (We will violate this golden rule for your small group project to save you time. Do as I say here, not as I am doing for you on that assignment.) Go to more than one source of information.
about the theory because there may be some differences in definitions even within the same theoretical perspective. Write down the definition you use. Include a one or two sentence definition of the construct and a description of the major differences in how the construct is defined and the decisions you had to make. For example, you might find that one theorist argues that there are three dimensions and another four. You have to make a decision about which to use.

**Identify variables that will represent the constructs**

You may find useful information in the literature that will help with this, but remember that the variables are specific to your study topic and the context (place, time, population) in which you will conduct the study. Make sure the variables represent all dimensions in the constructs of interest for your research. It is better to have too many variables than too few early in the process. You may want to generate measures for two or three variables representing the same construct because you can often merge the scores later if you decide to reduce the number of variables. You can also use this redundancy in the assessment of the validity of your measurements because multiple variables representing the same construct should have similar patterns of response. If you have multiple variables per construct, you can assess this convergence in scores. Multiple variables per construct are highly recommended.

**Item Development**

It is always valuable to look for existing instruments and use them if they are appropriate to the context for your study. Using an existing instrument – if it will work for you – is always preferable to creating your own. It takes a LOT of work to create an instrument. You must purchase or get the consent of the creator of an existing instrument to use it in your research. Failure to do so constitutes plagiarism. You can also sometimes “borrow” individual items from instruments others have developed. However, you need to cite the original work when you do this. If you are going to use more than one or two items, you need to get permission from the original creator to use the items. Even when whole instruments or items are available that seem like they will work for you, you need to test them thoroughly because context is critical to wording, content, and even response formats for items. In this class, I want you to practice developing your own instruments and your own items. Therefore, I expect you to rely primarily on items that you develop. Do not use existing instruments. No more than 25% of the items in any of the instruments you develop should be from the literature – 75% your ideas.

**Start with many redundant items.**

You will run multiple tests on your item banks to determine which items yield reliable, valid, and discriminatory results. Redundant means that you start with many items for each variable, items that differ significantly so that they capture the full meaning of the construct. You will eliminate most of them. You may also have two, three or more versions of an item that differ in wording or construction (for example, a reverse-scored version of an item) where such differences could influence how the respondent understands what you are asking. However, focus on redundancy in content, not wording. Taking an example of an instrument using closed responses (check the box), if I start with 30 items, I find that only 8 or 10 or them remain after all this testing. My rule of thumb is to start with 30-40 items per variable for instruments like scales or tests. For other instruments like an interview protocol with an open, narrative response, the key is to start with many different kinds of questions for each topic. For example, a leading question might start the conversation about a topic, followed by some probing questions (confirming and disconfirming) and end with some summary questions. We will look at the range of types of questions you can use in our exercise to develop an interview protocol. You would have several of the trails of questions for each variable just as you would have several items for a “check the box” set of responses. The principle is the same. **Start with many items in all assignments.**
Revise to correct problems in wording, response structure, etc. (technical issues)

Do not waste other people’s time looking at your instrument until you have done your best to correct technical problems. Bryman focuses a lot on these issues and provides excellent guidance. Fowler provides guidance as well. We also have resources about item structure and format at the course website. Do this before anyone else sees your instrument.

Expert Panel Review

After you have completed your review and correction, you need to get an expert panel to review your work. An expert panel consists of people who have expertise either in methods of data collection or in the topic of your research and/or your theoretical approach. This is not a check of wording, etc. Experts may find some of those problems, but it is your responsibility to fix those problems. The expert panel normally consists of your colleagues and for students the chair and members of your supervisory committee. Expert panel review is what your committee does when you defend your research proposal. In this class, you will largely use each other as members of the expert review panel. The work of the expert panel is largely conceptual. Typically, you would ask them to do three things for you. (1) The panel should provide an assessment of whether your instrument(s) capture the full meaning of the construct. (2) They should identify any aspects of your instrument that are not relevant to the theoretical construct or that are not appropriate for the topic/context of your study. One student and I struggled through several iterations, for example, to decide how to measure “young women’s life aspirations.” We both added and eliminated items and even variables in the process. (3) They should also assess the adequacy of the format and structure of the items. The common use of the “Likert-type” scale provides an example. People use it without thinking about whether they should use this approach or not. It is not appropriate for all contexts or topics. I once saw a Likert-type instrument with statements like “My department provides training for teaching assistants.” This is a simple “yes/no” item. There is nothing to agree or not agree with – it is a correct statement of fact or it is not. The Likert-like format of agree to disagree creates a great deal of mental work for the respondent (read a statement, think about it, decide if I agree with it or not, decide how much I agree with it). This mental work is valuable when topics are emotionally charged because it gives the respondent a way to establish distance from the potentially disturbing ideas expressed in the statements. If the topic is not one that is apt to create emotional turmoil for the respondent, do not use the Likert-type statement and agree/disagree response format. It is a waste of their time. You cannot use the Likert-type response format for this class unless there are sound reasons why some more straightforward approach will not work. Further, in general, providing people with statements rather than questions creates extra mental (cognitive) work and you will be penalized for using statements rather than straightforward questions.

Provide the panel members (again, that may be just one or two people) with the research questions, a brief description of your theoretical approach, and a full definition of each construct you are trying to operationalize. They do NOT need to see the standard items for demographics and such. It is also useful to draw their attention to areas where you are uncertain and feel you need their input. Do everything you can to make this as easy as possible for them to help you. Unless otherwise instructed, always conduct a cognitive review with members of this class, faculty members, or other graduate students with the requisite expertise for all assignments in this class.

Cognitive Testing

I regard cognitive testing as nearly equal in importance to expert panel review in the process of instrument development. Like expert panel review, you can reap major improvements from the process at relatively little expenditure of time and effort. It is a good use of your time. You can and should use this technique with all types of data collection. Cognitive testing must be done with members of the
target population for the study or with individuals who are “very much like” the members of the target population with regard to characteristics or traits that can affect how well they can respond to your questions. Several reading for the week we discuss operationalizing constructs discuss the use of cognitive testing (Collins, 2003; Castillo-Diaz & Padilla, 2013; Priede & Farrall, 2011). I expect you to use cognitive testing in virtually every assignment in this class.

Testing Instruments

The document “Some Useful Procedures for Operationalization” provides a specific discussion of various techniques for testing instruments. It covers techniques that I felt you needed to know about and use – but I could not find a good general reading about them appropriate for this class. In addition to this cheat sheet of mine, we will discuss additional techniques as we cover specific methods of data collection during the semester. The website for the week we discuss operationalization also includes many useful references, and so do the websites for the weeks when we focus on a specific method of data collection.

References


