Approach to Measurement
M. E. Swisher, Revised January 2020

This document describes the approach to measurement we will use in this course. It is divided into three discussions: (1) The terminology of research methodology and (2) operationalization. These are fundamental, core concepts. ***I assess your assignments based on your understanding and ability to apply these ideas.***

There is another document, Procedures for Instrument Development, that describes the specific steps you need to take and explains how to conduct the various tests you will perform as you create instruments.

Warning: Avoid the Qualitative Versus Quantitative Debate

In this course, we will apply the same standards and procedures to create research instruments for collecting data that provide numbers (quantitative) versus other types of scores like themes or categories (qualitative). Focus on the key concepts we employ – they are equally applicable to ANY form of data collection or analysis. Many contemporary methodologists, including me, find the quantitative versus qualitative discussion a distraction that does little to improve research. As a scientific realist, I draw no distinctions between the value of data created through any method and data in any form (numbers, word, categories, etc.). Your task in this class is to demonstrate that you understand the principles we study and can apply them to your work.

The Terminology of Research Methodology

Constructs

- **Constructs, the building blocks of theories, are typically broad concepts like self-actualization, religiosity, or identity.** Constructs have a theoretical meaning. Put another way, without theory, constructs (in science) are undefined. The scientific, theoretical definition differs from the everyday use of terms in most cases.

- ***Different theories often use the same term to refer to constructs that they define quite differently.*** 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.

- **Constructs and how they are defined are universal to a 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.

- **Most social scientific constructs have more than one dimension.** Socio-economic status is a good example. There are at least two dimensions in this construct – social status and economic status. The terminology can be confusing. 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, 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. In TPB attitude does **not** mean a "feeling or emotion." It really means a ***mental assessment*** of (1) what will happen if you engage in a behavior based that you believe will result in an outcome you want and (2) measuring the importance of this outcome to you. In other theories, attitude does mean a feeling, state of mind, mental state, or emotional state.

- **Precise definitions of the constructs that form the basis of scientific study are critical to reaching valid, reliable conclusions.** Step one in developing the methods for your research is to define the constructs you will explore fully and write the definitions the methodology section of the research proposal.

Variables

- **Unlike constructs, 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***.

- **Variables may be manifest or latent.** Manifest (or observable) variables are things that we can measure directly – simply by observing people, groups, organizations or any other unit of analysis. For social scientists, observing also includes factual information that people can give us like age or occupation. These variables are descriptive in nature. Most of our research relies on ***latent variables***, things that we cannot observe directly and that are not factual information that people or groups can provide to us. These are the variables that represent constructs. We create these variables in most cases by making a number of observations (which can mean a number of questions) that are relatively observable and then combining all of these observations to create a single variable.

- **Nevertheless, all of the variables used to represent one 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.

- **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 will have as many variables as there are dimensions in a construct. To measure attitude about losing weight using TPB, 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.

- **You may want to use several variables to represent a single construct or a single dimension of a construct for additional reasons and there is no one-to-one correspondence between**

variables and constructs or even between variables and dimensions of constructs. For example, 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. You can always combine scores in both qualitative and quantitative analyses. Therefore, 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, but often not. Creating the right items to ask about something like "power in the household" is not easy.

- **Few social scientific constructs can be well represented by single-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 surveys or polls conducted prior to an election use single items variables – like "which of the following four candidates do you plan to vote for?" Questionnaires also often use single item variables like age, gender, race, and educational level to describe the sample or population of interest. We will not spend any time on these simple variables in this course because they are easy to create, often standardized for things like household income categories, and typically are not explanatory or theory-based.

- **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. We need to create latent variables, and that almost always requires multiple observations.
  - 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 (discussed below) require multi-item variables.

The Terminology Mess and What We Will Use

The terminology is very confusing. Many authors never refer to indicators or items, but rather use the word variable for all three – treating each item as a variable or simply failing to distinguish between the two. Be careful in your use of these terms and **please use the definitions provided here for this class.**

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.

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. But even here things get messy. Income is an example. It can mean just the *salary or wages* a person earns or it can mean the *earnings from all sources*. ***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.

Operationalization

Operationalization is the process of going from an abstract theoretical concept or construct to a list of questions or items that people can answer.
Bad process = bad information = meaningless results

The challenge for social science researchers is that we must turn an abstract idea, a theoretical construct like socio-economic status or resilience, into questions that people can answer. Asking someone: "On a scale of 1 to 10, what's your socioeconomic status?" will not produce a meaningful answer. Even if we are ask an individual to describe his/her socioeconomic status, most of us could not answer the question well. Our task is to create a set of items in the form of questions, statements, or observations that make sense to respondents – that they can answer relatively easily. We focus on three ways of making sure we do not produce meaningless results: reliability, validity and discriminatory power.

Reliability of Measures

There are two aspects to reliability, stability and consistency of responses.

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 administer the same set of items to the same people on two occasions.

Consistency **means that each individual responds similarly to items that are supposed to measure the same construct.** We expect different individuals to give different responses to any question, but *any one individual* should give similar answers if asked two or more related or similar questions. Here is an example.

I ask you which of five issues (poverty, the economy, defense, social justice and the environment) is most important to you. You respond "the environment." I later ask you to rank five proposed uses of funds (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 two items (the issue and the way to use tax money) represent the same constructs, you should rank

"fund research to combat climate change" highest when you respond to the question about funding. 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. This is an open response format, unlike the closed response formats for the first two questions. You might respond "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" in qualitative data analysis. In this example, you have given three responses to three related items that are consistent. In essence, I asked how important different goals of government are to you in three different ways, and every time you gave me the same answer – the environment. We will use techniques like Cronbach's alpha and inter-rater reliability to assess reliability.

Reliability and validity are inter-dependent. 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 value indicates that people respond in a similar pattern to a set of items. That suggests 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

Validity in science means making sure that your measurements (items) captures the meaning of the theoretical constructs of interest – as you defined them – and are appropriate in the context of your research. Consider three questions when addressing validity.

- Did you measure what you needed to measure to answer your research question?
- Do the measures and scores accurately capture the 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)?

We will divide validity into two broad concepts – face validity and measurement validity. The discussion of validity is confusing, in part based on epistemological differences. Different authors use the same term to mean different things. Some refer to many different forms of validity like content validity, criterion validity, face validity, and construct validity, while others make no distinctions. Authors often do not agree on the definitions of these terms. **I want you to focus on just the two broad aspects of validity – face validity and measurement. Do NOT use the checklists of a litany of types of validity.** However, I want you to be able to use the research methodology literature that discusses other ideas about validity because you will see these ideas in the literature. Therefore, I provide resources that use other terminology and other definitions of validity in social scientific research. You can compare approaches to validity, but I am not interested in your ability to repeat what others say. I encourage you to think about validity in a sophisticated way – not by using a "checklist" approach. ***Please discuss both aspects (face and measurement) of validity in your responses in assignments. I often get submissions that focus solely on face validity. That is not acceptable.***

Face validity is the easiest to understand, but there are few explicit guidelines 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 pose capture the actual *experiences* of people. This is the overlap between the world of phenomena or experience and the world of words and ideas as shown in the diagram on p. 18 in Watt and van den Berg. 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). Ask yourself and the people who review your instruments: "On the face of it, does this item make sense in terms of its relevance to

the constructs of interest in my study ***and*** people's ability to understand and respond to it?" Many researchers treat face validity as separate from content validity. I prefer the approach of Watt and van den Berg who treat face validity as distinct from measurement validity. Here is an example of the difficulty in determining face validity.

> I want to understand how ***food insecurity*** affects people's ***self-respect and self-confidence***. I cannot simply ask: "On a scale of 1 to 10, how food insecure is your family?" People cannot answer that question. I have to ascertain what food insecurity *how people experience food insecurity*. If can ask "do you get food stamps," but most experts will say that this single item 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 for groceries. The question "do you eat out at least once a week" is not a good item either. We hear many negative comments about how poor people could save money and eat better if they would just not eat fast food. Food insecure people may do so because they hold two jobs, have many other obligations, and do not have time to cook. Hence, eating out at the local fast food joint may have nothing to do with food security status for many people. Food insecurity, self-respect and self-confidence require careful operationalization to answer a question about the relationships of interest.

You must also consider the ***context for a study when you think about face validity***. The meaning of concepts like food insecurity and income differ among nations and within individual countries. For example, if you want to know about income, you have to make sure your questions are relevant for the social and economic context. 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 thinking about the context.

Two groups of people can provide initial insights about the face validity of research instruments – experts in social research methods and experts in the topic of the research. Get input from topical experts first to determine whether experts in **the topic of your research** agree that your items capture the needed content to answer your research question(s). Once you have some agreement that the items "make sense theoretically," consult with methodological experts to get advice about response formats, order in the instrument, clarity of instructions, use of various elicitation techniques, and wording. Do **not focus on wording alone**. The methodological review is a consideration of the overall structure of the instrument, not just whether you used the right words.

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." We will address this primarily through cognitive testing techniques. This group consists of people who are members of the target population **or** are very similar to members of the target population with regard to characteristics that could affect how they would understand and respond to the instrument. 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."

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 they are defined theoretically and procedures we use to ensure that individual measurements 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, 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 measurement*. I expect you to use these techniques in this class.

Divergent or discriminant validity **refers to the degree to which measures of *unrelated or different constructs* produce different scores.** It is the opposite of congruent validity. You can assess divergent validity in two ways: The degree to which scores for *different constructs vary in your instruments* and 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 among individuals.

Construct validity **adds two additional requirements for measurement validity, 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.*

- Define the constructs clearly and precisely with items must capture ***all*** of the components in the definition.
- 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 is not true in the sense that the predicted match is weak or even absent in the empirical data (see Bhattacherjee, 2012). This is why it is so important to define the dimensions and *create a distinct variable score for each one.*
- ***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 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 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 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.

References

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.

Liu, L., Li, C. & Zhu, D. (2012) A new approach to testing nomological validity and its application to a second-order measurement model of trust. *Journal of the Association for Information Systems* 13(12), 950-975.

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).