e-Source
Behavioral & Social Sciences Research

Objective Measurement of Subjective Phenomena

1. Learning Objectives

After reviewing this chapter readers should be able to:

- Define and understand the basic elements of measuring behavioral outcomes.
- Identify different types of behavioral outcomes and the measurement procedures for assessing them.
- List and give examples of methods of constructing measures, along with the problems and biases that may arise when assessing constructs.
- Identify and define different types of reliability, distinguishing among types of reliability and their unique insights into the assessment of outcomes.
- Define traditional forms of validity – content, criterion-related, and construct validity – and understand how convergent and discriminant validity offers clearer information regarding validity.

Office of Behavioral and
Social Sciences Research

Department of Health
and Human Services

National Institutes
of Health

2. Introduction

Assigning numbers to individuals to represent the magnitude or presence vs. absence of an attribute or characteristic (Allen & Yen, 1979; McDonald, 1999).

When we measure a human characteristic well, we gain a valuable description of individuals on the dimension of interest. However, in the behavioral and social sciences, we often intend to measure dimensions – such as anxiety, loneliness, or social support – that are intrinsically difficult to measure, especially when compared with measurement of corporeal dimensions, such as blood pressure, glucose levels, or height and weight. Despite difficulties arising when measuring dimensions like anxiety and loneliness, accurate measurement is a valuable adjunct in many everyday treatment situations and is the backbone of basic and applied research in science.

2. Introduction

Individual differences

A most striking aspect of humans is the presence of individual differences in most personal characteristics. Some personal characteristics lead to groupings of persons, such as ethnic status (European American, African American, etc.). Other characteristics lead to individual differences that fall on a continuum, much as height (varying continuously from short to tall). Further, these differences can be:

- Interindividual differences (differences between individuals on a given dimension); or
- Intraindividual differences (differences within individuals), such as different levels of anxiety in a single individual as a function of context.

How do we capture or assess these individual differences?

As described by McDonald (1999), we can use:

Informal Characterizations

Construing or capturing individual differences using non-standardized forms of assessment, such as verbal descriptions of a person from self-reported stories, observations of others, or works of literature.

Semi-Formal Approaches

Open-ended interviews of participants in which a few standard probes are provided to initiate the interview and the interview proceeds from that point.

Formal Systems

Consistent and precise measures administered to every person.

Formal systems of measurement have all or most of the following four characteristics:

- **Standard measurement operations**: a prescribed way of delivering the assessment, including context (e.g., individual vs. group administration) and form (e.g., paper-and-pencil forms vs. computerized administration)
- **Standard set of items**: a defined set of items that is administered to all persons

- **Specified forms of manifest (observed) scores**: a specific way of combining information from items / indicators to obtain raw scores for individuals
- **Ways of standardizing scores**: a formulaic way to obtain standardized scores so that the score for an individual can be interpreted relative to some norming population

Formal Requirements for Measurement

Recall that measurement involves assigning numbers to individuals to represent the magnitude or presence vs. absence of an attribute for each person. Given this goal, we need the following requirements:

- **Requirement 1**: a clear description of the attribute or characteristic to be assessed
- **Requirement 2**: a scheme of numbers
- **Requirement 3**: an operational tie between numbers and the magnitude or the presence vs. absence of the attribute
- **Requirement 4**: a standard way of assigning numbers to individuals to reflect the magnitude or presence of the attribute

3. The Construct, or Characteristic, to be Measured

Ease of Measurement

Constructs vary in their ease of measurement, with some constructs being relatively easy to assess and others requiring more subtle or indirect measurement.

Direct:

Some attributes or constructs can be measured directly. In medical settings, direct measurements are often obtained on routine doctor visits.

When measuring behavioral outcomes in the social sciences, the personal characteristic to be assessed is called a **construct** (Cronbach & Meehl, 1955; Messick, 1995). The construct is a proposed attribute of a person that often cannot be measured directly, but can be assessed using a number of indicators or manifest variables.

Constructs are also discussed under other labels, such as theoretical constructs or latent variables, which are interchangeable terms.

Example 1

Direct construct examples:

1. Height (in inches or cm)
2. Weight (in lbs or kg)
3. Blood pressure (in mmHg)

Indirect:

In the behavioral and social sciences, we usually must use more indirect ways to measure constructs, so we develop a number of items to assess the construct.

Example 2

Indirect construct examples:

1. **Depression** - Scales for depression often consist of 10 to 20 items or more, and the score for depression is a sum of scores on these items.
2. **Happiness** - Happiness is a narrower construct than depression, but a happiness scale might still require 5 to 10 items or more to assess well.

Note: Ease or directness of measurement is not an indicator of how closely related a scale score is to an underlying construct or how important the attribute is for a given problem.

3. The Construct, or Characteristic, to be Measured

Theoretical Requirements

The construct or attribute must be carefully defined and delineated (Jackson, 1971). Theory regarding an attribute involves matters such as whether the construct is:

- dynamic, fluctuating over time, or stable across time;
- dependent on context or not; and
- occurs in only some individuals or in all individuals.

Answers to such questions are an invaluable aid in deciding how to measure an attribute.

Empirical Requirements

Prior research provides a valuable context for work on measuring a construct (Cronbach & Meehl, 1955; Campbell & Fiske, 1959). If prior attempts to assess the same construct have met with some success, then current efforts can be informed by these successes. Or, prior attempts to assess a similar construct may have consistently failed to yield expected results. Such information would still be quite valuable for work to develop a new measure of a construct, as it may indicate the need to strike off on a different path to measuring the attribute.

4. Nature of the Construct

Personal characteristics differ in the nature of individual differences that are presumed to exist. As a result, the researcher must outline the nature of the personal characteristic to be measured. When measuring a characteristic, one might consider the following dimensions:

Dimension 1: Form of individual differences to be exhibited

Individual differences on an attribute of interest may be quantitative or may be qualitative. Quantitative differences are typically seen indexing “more vs. less” of an attribute along a continuous scale, whereas qualitative differences usually take the form of identifying either a group of which the person is a member or a distinct characteristic that a person possesses (or does not possess) (Waller & Meehl, 1998; Widiger & Trull, 2007).

Continuous distribution:

A continuous distribution is a very common conception, in which individual differences are represented by numbers on a scale that indicates a person has more (or less) of the characteristic.

Example 3

Continuous behavioral outcome examples:

1. **Intelligence** - As assessed using an individually administered intelligence test and indexed by the intelligence quotient (IQ). IQs are usually normed to have a mean of 100 and SD of 15 in the population, and IQs are reported as whole numbers.
2. **Extraversion** - Which is often assessed using 10 to 20 items, each answered on a 1-to-5 or 1-to-7 scale. Summing across items results in scale scores, with higher scores indicating higher levels of extraversion.

Dichotomous Distribution:

One version of a categorical scale, a dichotomous distribution indicates whether a person falls in one or the other of two mutually exclusive and exhaustive classes or groups. Thus, a dichotomous distribution involves making a binary choice of group membership for each person.

Example 4

Dichotomous behavioral outcome examples:

1. **Clinical depression** - Here, one would decide whether a person meets diagnostic criteria of clinical depression by exhibiting a sufficient number of signs or symptoms of depression.
2. **Mental retardation** - A person must meet three criteria – low intelligence, deficits in adaptive behavior, and appearance of these criteria prior to the age of 18 years – to be diagnosed with mental retardation (which is now called intellectual disability).

Polytomous Distribution:

A polytomous distribution is another version of categorical measurement whereby individuals are sorted into more than two mutually exclusive and exhaustive categories.

Example 5

Polytomous behavioral outcome example:

Attention Deficit/Hyperactivity Disorder (ADHD). ADHD is often identified using one set of symptoms for attention deficits and another set for hyperactivity. Then, a child might fall into one of four groups:

1 = no ADHD
2 = ADHD, attention deficit alone
3 = ADHD, hyperactivity alone
4 = ADHD, combined attention deficit and hyperactivity

Ordered Categorical Scale:

An ordered categorical scale is one on which numbers indicate more or less of an attribute, but score intervals are not equal. Thus, scale scores seem similar to those on a continuous scale, but scores on an ordered categorical scale do not fall on an equal-interval scale. Most rating scales used in the social and behavioral sciences are most accurately characterized as falling on ordered categorical scales.

Example 6

Ordered categorical scale example:

Questions on many self-report inventories ask respondents to indicate their response to each item on a 1-to-5 scale, ranging from 1 = strongly disagree to 5 = strongly agree. Without a substantial amount of work, it is difficult to justify the assertion that the difference between scores of 1 and 2 is equal to the difference between scores of 3 and 4.

4. Nature of the Construct

Dimension 2: Breadth vs. narrowness of the construct

Constructs vary considerably in their breadth. Some constructs are very broad and subsume considerable variation in content, whereas other constructs are much narrower in the content subsumed. This dimension is often discussed under the rubric of “bandwidth vs. fidelity” (Clark & Watson, 1995).

Broad Constructs

Broad constructs are those that cover a wide range of behavioral exemplars, meaning that assessment of a broad construct should be based on sampling from several subdomains of content.

Example 7

Broad construct examples:

- General intelligence - Represented by the Full Scale IQ from an intelligence test, which should be based on multiple kinds of cognitive function; and
- Extraversion - Has a number of facets, including talkativeness or gregariousness, assertiveness in social situations, and activity level.

Narrower constructs

Narrower constructs cover a much narrower range of behavioral content.

Example 8

Narrower construct examples:

- Numerical facility - A subset of the domain of intelligence, which refers to speed and accuracy of responding to simple arithmetic problems, such as addition and subtraction; and
- Gregariousness or assertiveness - Are two subdomains of extraversion.

4. Nature of the Construct

Dimension 3: Context dependence

Some constructs are thought to be relatively independent of context, whereas others seem to be much more dependent on or affected by context (Donnellan, Lucas, & Fleeson, 2009; Lucas & Donnellan, 2009).

Example 9

Context-independent construct examples:

1. **Chronic depression** – A person suffering from chronic depression will typically exhibit signs and symptoms of depression regardless of surroundings.
2. **General intelligence** – A person with high intelligence tends to exhibit greater facility with a wide range of intellectual problems and issues than does a person of low intelligence.

Example 10

Context-dependent construct examples:

1. **Certain phobias** – These are relatively context-dependent. For example, agoraphobia is fear of a panic attack in a situation offering few easy means of escape, such as a new, open area.
2. **Test anxiety** – Test anxiety is a form of anxiety that arises in situations in which a person feels symptoms of anxiety only surrounding examinations of their performance.

4. Nature of the Construct

Dimension 4: Temporal constancy (or consistency or stability) versus fluctuation (or instability)

The dimension of temporal constancy can be used to distinguish trait construct, which are stable over time, from state constructs, which fluctuate notably over time (Gaudry, Vagg, & Spielberger, 1975; Hampson & Goldberg, 2006).

Example 11

Trait construct examples:

1. **Trait anxiety** – This is indexed by asking a person how s/he has felt, in general, over an extended period of time, such as the last month or last six months.
2. **Big 5 dimensions of personality** – These are thought to be relatively stable descriptions of an individual. They include:
   - Extraversion
   - Agreeableness
   - Conscientiousness
   - Neuroticism
   - Openness to Experience

Example 12

State construct examples:

1. State anxiety – State anxiety is assessed by asking a person to report feelings of fear, uneasiness, or shortness of breath “right now” or “today.”
2. Bipolar disorder – Bipolar disorder is characterized between swings between more or less manic behaviors over time.

4. Nature of the Construct

Dimension 5: Temporal duration

An alternative way of characterizing the temporal dimension is the temporal duration of the characteristic. Acute problems are those that may be marked at the present time, but are expected to wane over time, whereas chronic problems are those likely to remain invariant over time or to recur predictably across time.

Example 13

Acute problem examples:

1. Panic attack - A panic attack can be extremely strong and florid at a given time, but may wane rather rapidly and recur only intermittently.
2. Major depressive episode - A major depressive episode can be a response to a major life event or series of event (e.g., death of a significant other, loss of job) and may not recur.

Example 14

Chronic problem example:

Autism - Autism is a blanket term for a spectrum of problems related to language and communication, social functioning, and (often) repetitive behaviors. Although some children with autism appear to improve notably across time, autistic behaviors tend to be problems difficult to remediate.

4. Nature of the Construct

Dimension 6: Developmental course

The developmental course of many behaviors involves both growth, development, and regulation during the early years of life and aging declines or disintegration during the later stages of life (Horn & Hofer, 1992; Soto, John, Gosling, & Potter, 2008; Srivastava, John, Gosling, & Potter, 2003). Examples of each are given below.

Example 15

Growth examples:

1. **Height from infancy through early adulthood** - After fairly steady increases in height, most adolescents show a rapid growth spurt closely associated with puberty, after which growth slows and is usually complete by early adulthood.
2. **Mental age** - The concept of mental age presumes that intelligence increases steadily with age during the developmental period.

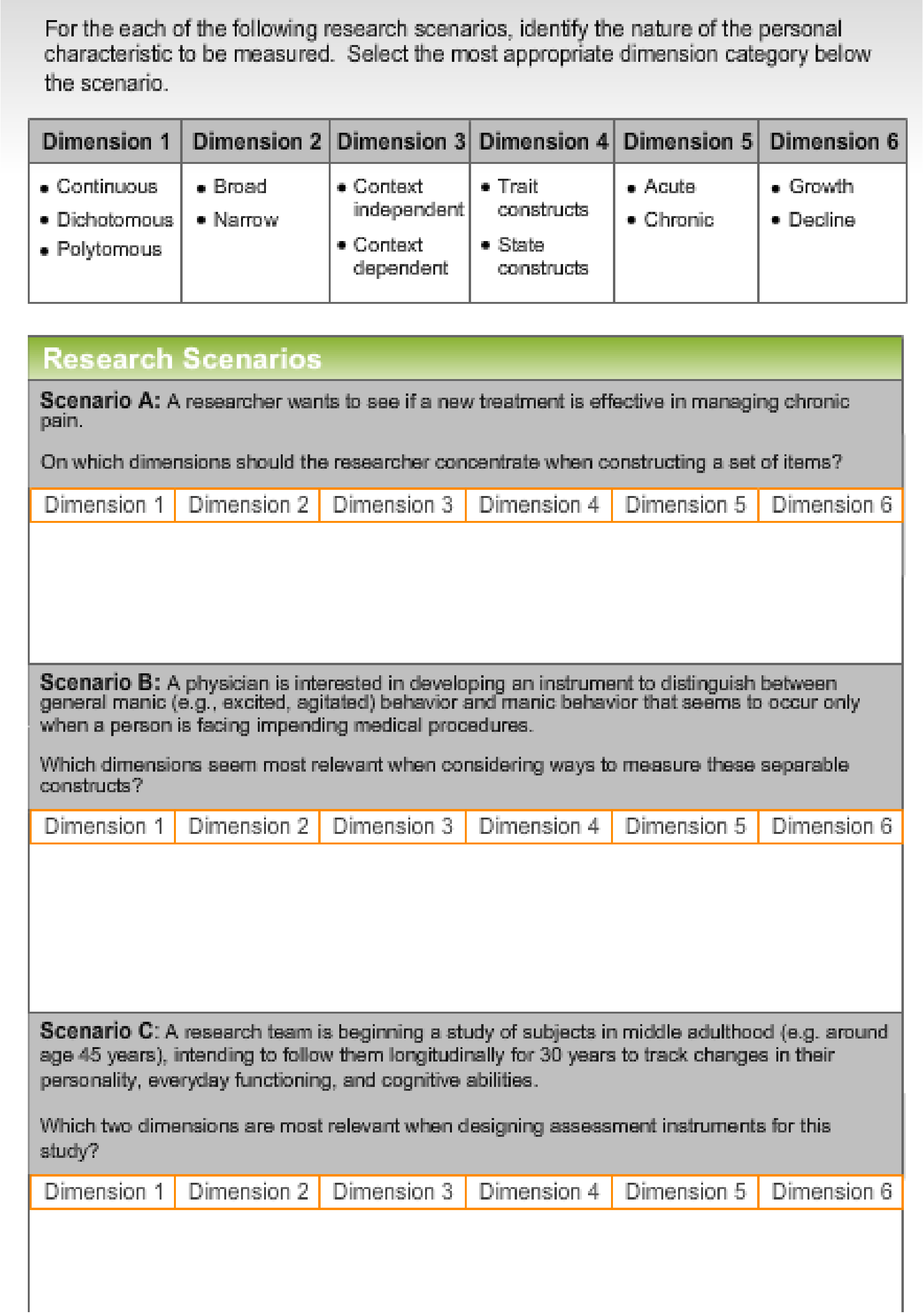
Example 16

Decline examples:

1. **Memory performance** - On both long-term memory and short-term memory tasks, adults tend to show systematic declines in performance after the age of 40 or 50 years.
2. **Speed of response** - Speed of response tends to decline sooner that most other mental skills, declining notably and systematically after age 30.

4. Nature of the Construct

Exercise 1



For the each of the following research scenarios, identify the nature of the personal characteristic to be measured. Select the most appropriate dimension category below the scenario.

Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6
• Continuous • Dichotomous • Polytomous	• Broad • Narrow	• Context independent • Context dependent	• Trait constructs • State constructs	• Acute • Chronic	• Growth • Decline

Research Scenarios

Scenario A: A researcher wants to see if a new treatment is effective in managing chronic pain.

On which dimensions should the researcher concentrate when constructing a set of items?

Dimension 1 | Dimension 2 | Dimension 3 | Dimension 4 | Dimension 5 | Dimension 6

Scenario B: A physician is interested in developing an instrument to distinguish between general manic (e.g., excited, agitated) behavior and manic behavior that seems to occur only when a person is facing impending medical procedures.

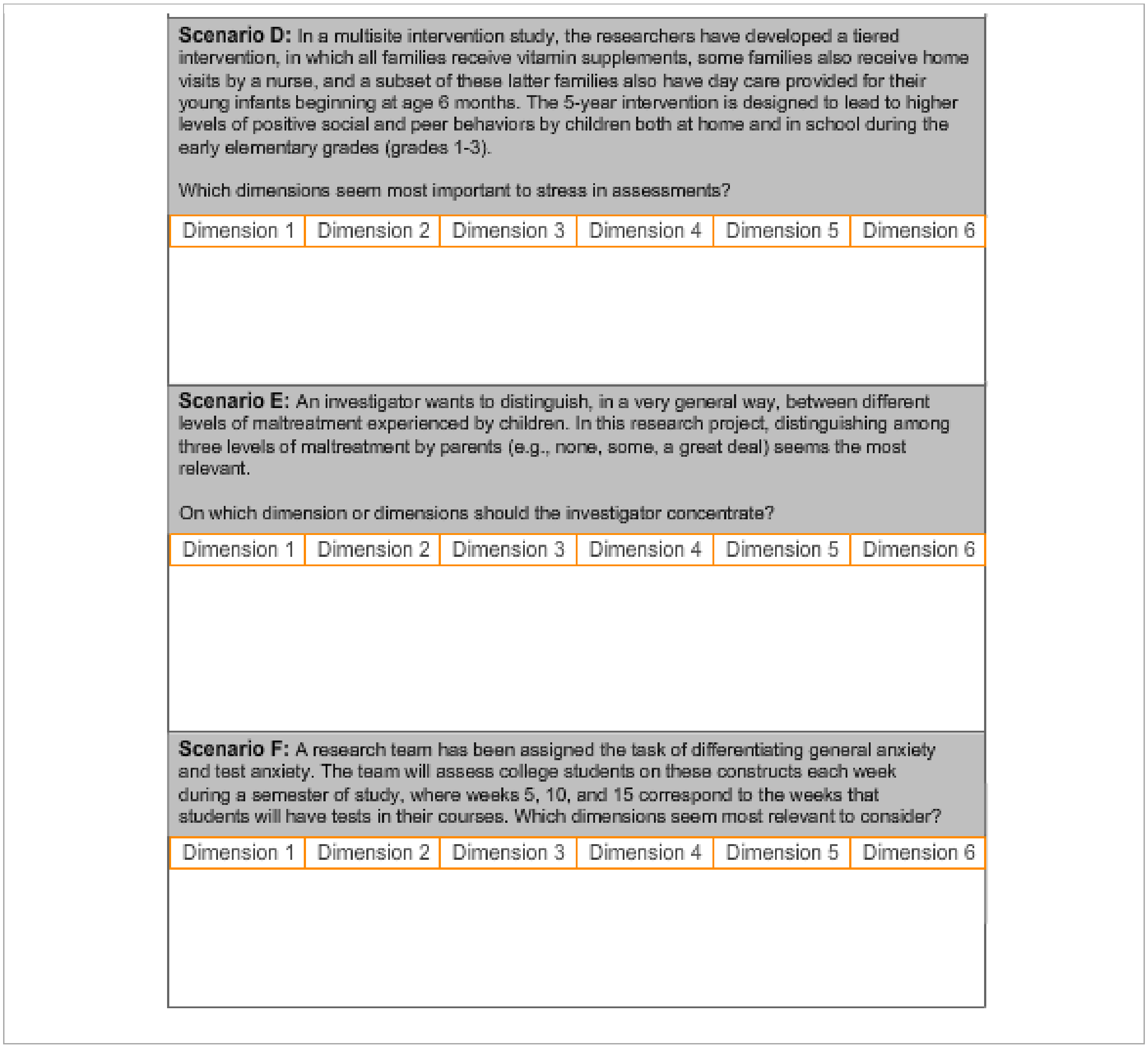
Which dimensions seem most relevant when considering ways to measure these separable constructs?

Dimension 1 | Dimension 2 | Dimension 3 | Dimension 4 | Dimension 5 | Dimension 6

Scenario C: A research team is beginning a study of subjects in middle adulthood (e.g. around age 45 years), intending to follow them longitudinally for 30 years to track changes in their personality, everyday functioning, and cognitive abilities.

Which two dimensions are most relevant when designing assessment instruments for this study?

Dimension 1 | Dimension 2 | Dimension 3 | Dimension 4 | Dimension 5 | Dimension 6



Scenario D: In a multisite intervention study, the researchers have developed a tiered intervention, in which all families receive vitamin supplements, some families also receive home visits by a nurse, and a subset of these latter families also have day care provided for their young infants beginning at age 6 months. The 5-year intervention is designed to lead to higher levels of positive social and peer behaviors by children both at home and in school during the early elementary grades (grades 1-3).

Which dimensions seem most important to stress in assessments?

Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6

Scenario E: An investigator wants to distinguish, in a very general way, between different levels of maltreatment experienced by children. In this research project, distinguishing among three levels of maltreatment by parents (e.g., none, some, a great deal) seems the most relevant.

On which dimension or dimensions should the investigator concentrate?

Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6

Scenario F: A research team has been assigned the task of differentiating general anxiety and test anxiety. The team will assess college students on these constructs each week during a semester of study, where weeks 5, 10, and 15 correspond to the weeks that students will have tests in their courses. Which dimensions seem most relevant to consider?

Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6

5. Items, Levels of Measurement, and Methods of Scale Construction

Items

Two general categories of items: Objective and non-objective items (McDonald, 1999). Objective items are those that involve no subjectivity when scoring responses. Conversely, non-objective (or subjective) items are items that leave some room for subjectivity in scoring. Given their preponderance in survey methodology, we concentrate here on objective items.

Types of objective items: Objective items come in many different forms, several of which are shown below (see McDonald, 1999), for a more extensive review of item types):

Completion items state a problem, and the respondent must generate an answer.

Example 17

Completion item example:

Example: 5 + 4 = ____

Multiple-choice items provide a question stem and several answer options; the test taker must select one (or more) of the options as the optimal answer.

Example 18

Multiple-choice item example:

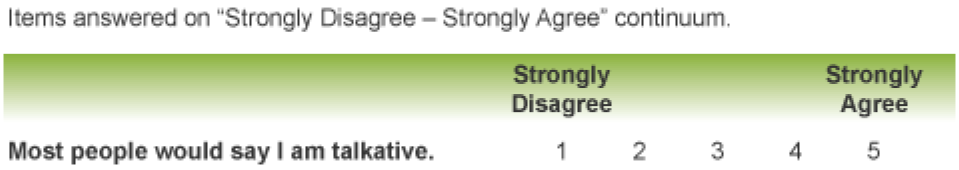
The mean of a distribution is a measure of

1. location
2. standard deviation
3. variance
4. range

Ordered-category items allow respondents to register their response on a graded continuum, which is a very common approach to measuring many behavioral outcomes.

Example 19

Ordered-category item example:

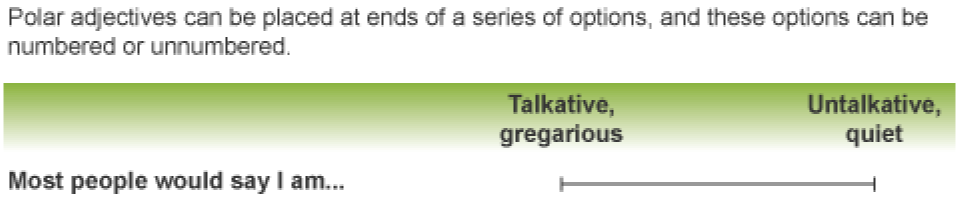


Items answered on "Strongly Disagree – Strongly Agree" continuum.

	Strongly Disagree				Strongly Agree
Most people would say I am talkative.	1	2	3	4	5

Example 20

Ordered-category item example:



Polar adjectives can be placed at ends of a series of options, and these options can be numbered or unnumbered.

	Talkative, gregarious	Untalkative, quiet
Most people would say I am...		

5. Items, Levels of Measurement, and Methods of Scale Construction

Item Scores and Test Scores

The number assigned to an item response is a code or score. We call the number a code if it distinguishes between two categories, and we call the number a score if we plan to perform numerical operations on the number (McDonald, 1999).

Depending on item type, items can be scored in binary or integer fashion. Binary scoring refers to 0-1 scores, such as scores of 0 = incorrect, 1 = correct. Integer scoring is used when providing scores using more than 2 points, such as scores varying from 1-to-7, which may reflect judgments on a continuum ranging from 1 = strongly disagree to 7 = strongly agree.

Item writing: Many helpful ideas about writing items clearly, formatting measurement options for items, and the general “nuts and bolts” of dealing with items can be found in general sources such as McDonald (1999) and Nunnally and Bernstein (1994).

Two or more items, when taken together, constitute a “test” or “scale.” The total score on the scale is typically intended to measure an underlying attribute or characteristic (i.e., construct). The scale score is formed by simply summing the item scores. If we divide a test into subtests for distinct attributes, we form subscales. Subscales can be formed on an a priori or theoretical basis or can be formed on an empirical basis, as discussed below (Allen & Yen, 1979; McDonald, 1999). A thorough discussion of how to construct scales is given in DeVellis (2003).

5. Items, Levels of Measurement, and Methods of Scale Construction

Levels of Measurement

Levels of measurement, a topic of concern for over 50 years, have been distinguished for at least two reasons. First, levels of measurement are schemes of numbers for representing attributes of persons, so these levels of measurement serve basic requirements of assessment. Second, the level of measurement for a given attribute may limit the kinds of statistical manipulations that can be conducted with the numbers, although this has been and remains a point of contention.

Researchers typically distinguish among four basic levels of measurement – including nominal, ordinal, interval, and ratio scales (McDonald, 1999; Nunnally & Bernstein, 1994).

LEVELS OF MEASUREMENT

Nominal (or Categorical) Level of Measurement

The number assigned to an individual indicates a class or group of which the person is a member. Using nominal measures, a researcher can distinguish between two or more than two classes for a particular attribute. The following examples illustrate nominal measurement:

1. **Religion**
    - 1 = Protestant
    - 2 = Catholic
    - 3 = Jewish
    - 4 = Islamic
    - 5 = other

2. **Clinically depressed**
    - 0 = no
    - 1 = yes

Ordinal Level of Measurement

An ordinal scale consists of a set of numbers varying along a continuum. A higher number indicates “more” of the attribute, so a higher number indicates a greater magnitude of the attribute, but the intervals are not equal in size. Consider a scale with numbers that vary from 1 to 9. On this scale, a value of 3 indicates more of the attribute than 2 or 1 and less of the attribute than a score of 4 or higher. But, if scores fall on an ordinal scale, the difference between 2 and 3 is not necessarily equal to the difference between 4 and 5 (or any other 1-point difference). Examples are:

1. **Mohs hardness scale**

   1 = talc
   2 = gypsum
   ...
   10 = diamond

2. **Typical item from a scale for marital satisfaction: I generally feel happy and satisfied with my marriage.**

   1 = strongly disagree
   2 = disagree
   3 = neither agree nor disagree
   4 = agree
   5 = strongly agree

Interval Level of Measurement

As with an ordinal scale, increasing numbers on an interval scale indicate “more” of an attribute so indicate greater magnitude. But, an interval scale adds the criterion of equal intervals, although the zero point on the scale is arbitrary. As a result, ratios of values on an interval scale cannot be interpreted meaningfully, but ratios of differences between scale values are meaningful. Examples include:

1. **Temperature in Fahrenheit degrees** - The Fahrenheit scale refers to a scale on which 32 degrees (or 32 °F) is the melting point of ice at sea level and 212 degrees is the boiling point of water at sea level, with equal intervals along the scale.

2. **Temperature in °C** - The Centigrade scale is similar to the Fahrenheit scale, but has

different scale points designed for the melting point and boiling point of water. The Centigrade scale also has equal intervals, but only 100 scale points between the melting and boiling points of water, so a 5 point change in °C is identical to the temperature change of 9 points in °F.

Ratio Level of Measurement

The ratio level of measurement is the same as interval level, but adds the criterion of an absolute (or rational) zero point. That is, the arbitrary zero point for an interval scale is replaced by a rational or absolute zero point on a ratio scale. Very few variables in the social and behavioral sciences fall on a ratio scale. Examples are:

1. **Temperature on the Kelvin scale** - The Kelvin scale (denoted using the letter K) uses the same scale intervals as the Centigrade scale, so a 5 point change on the Kelvin scale is equal to a 5 point change on the Centigrade scale. However, the zero point of the Kelvin scale is equal to approximately -273 on the Centigrade scale and represents the absence of thermal energy. Thus, it is accurate to say that a temperature of 10 K is twice as hot as a temperature of 5 K, even though a temperature of 10 °C is not twice as hot as 5 °C.

2. **Reaction time** - One behavioral scale that can be argued to fall on a ratio scale is reaction time. When studying cognitive processes, psychologists administer problems via computer and measure the time taken to respond to the problem. During aging, mental processes tend to slow considerably, and one would be justified in claiming that a reaction time of 1000 ms. is twice as long as a reaction time of 500 ms.

5. Items, Levels of Measurement, and Methods of Scale Construction

Methods of Scale Construction

Different methods or approaches of constructing scales or tests have been described over the past half-century. These different methods constitute alternate ways of analyzing items from a scale, retaining items that assess a construct well and deleting items that do not.

We often break down these methods into three general approaches, which go by the names **empirical keying, factor analytic, and rational** (Burisch, 1984; Hase & Goldberg, 1967).

To utilize any of these three methods, one should have access to a large number of items. Then, several steps are needed to implement scale development under a given approach.

Method 1: Empirical keying

1. Identify groups you want to differentiate, such as schizophrenic vs. non-schizophrenic (or “normal”), or high achievement drive vs. low achievement drive
2. Administer items to individuals identified as belonging in the groups
3. Perform analyses to identify items that persons in the opposed groups answer in significantly different ways (e.g., endorse at significantly different rates)
4. Best items: those that discriminate best between groups

Method 2: Factor analytic

1. Administer items to large sample(s) of individuals
2. Compute correlations among items
3. Use factor analysis to determine how many dimensions underlie the data – and interpret the factors that emerge
4. Best items: those that load highest on each factor

Method 3: Rational

1. Start with careful delineation of construct

2. Write items to capture the construct by combining behaviors that exemplify the construct in contexts within which those behaviors might be exhibited
3. Use targeted statistical methods (e.g., item analysis) to select best items
4. Best items: those that correlate highest with the sum of the rest of items for a scale

5. Items, Levels of Measurement, and Methods of Scale Construction

Comparisons of methods of scale construction

Sometimes one is attempting to assess a complex, multifaceted conception, as occurs in diagnostic settings that requires documentation of presence of several symptoms. This research approach would favor use of **Method 1** (Empirical keying).

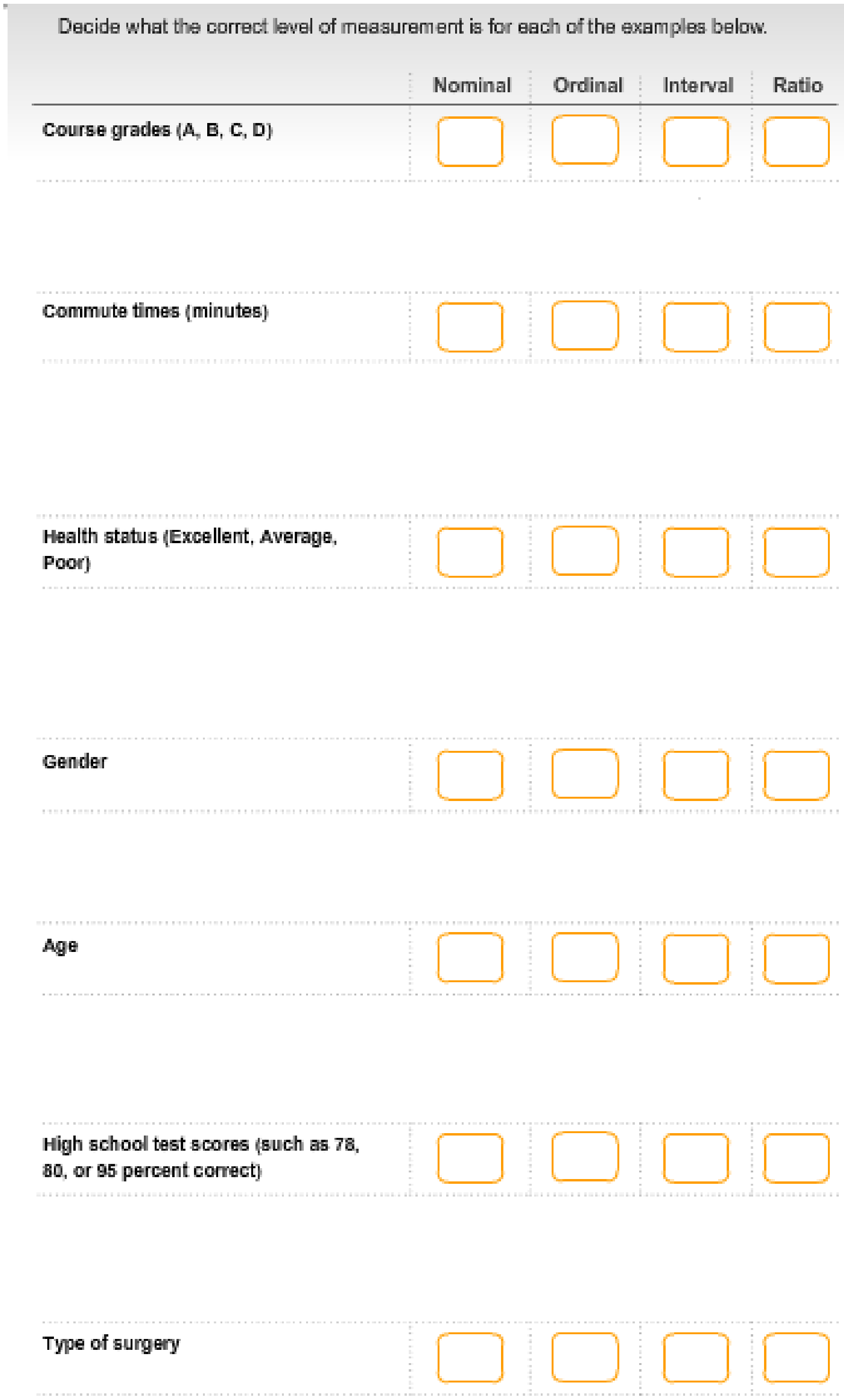
If one has a commitment to empirical validation and for the inductive derivation of theory from data, then **Method 2** (Factor analytic) might be optimal.

If the research aims concentration on development of scales to test a priori theories, then **Method 3** (Rational) is probably the method to choose.

> **Importantly, in empirical comparisons, all three methods arrive at scales that have comparable levels of reliability and validity (topics that will be discussed soon). Thus, if strong psychometric properties for a scale are the goal of a project, all three methods of scale construction can be recommended.**

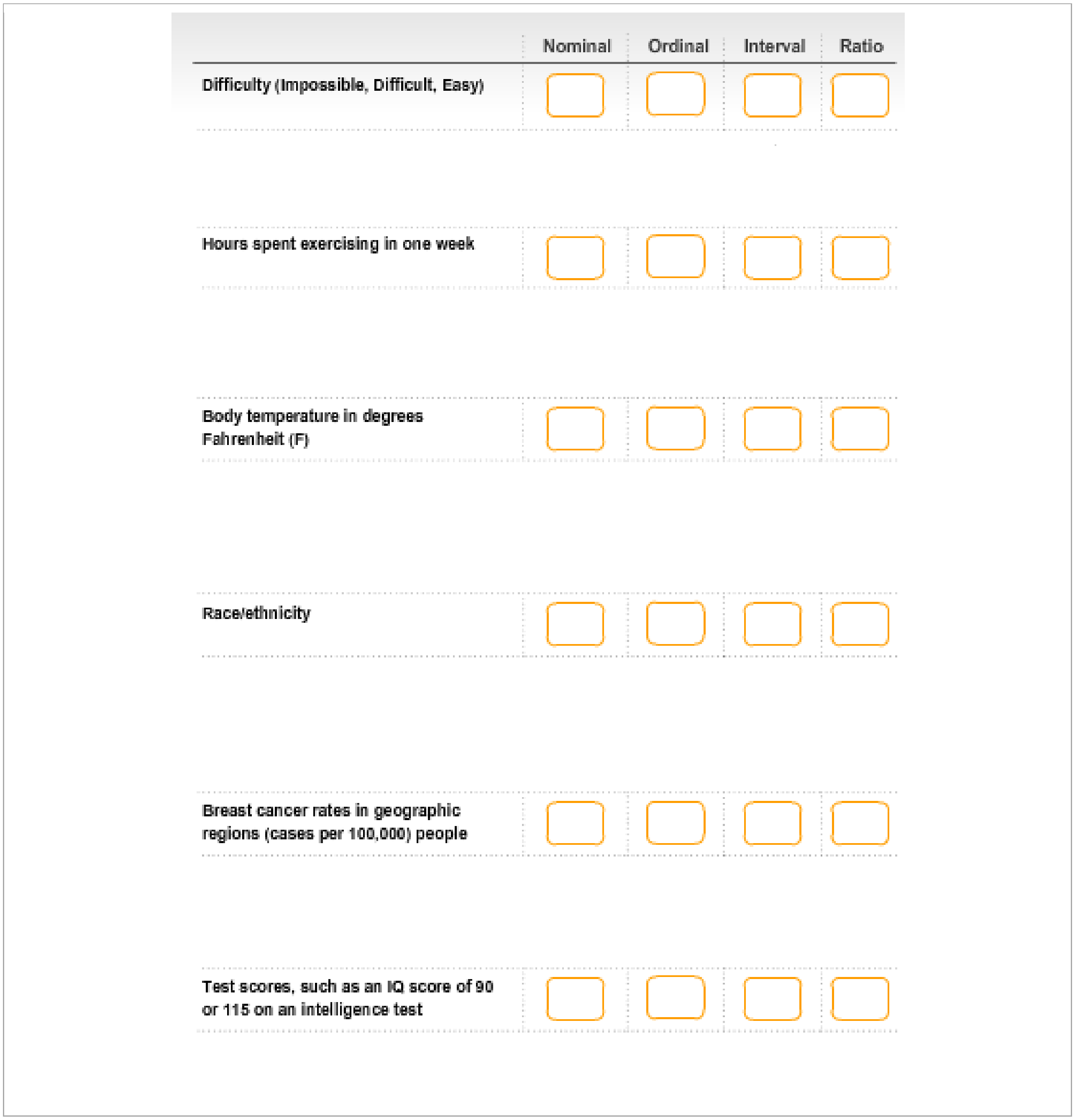
See Hase and Goldberg (1967) for an early extensive test of alternative methods of scale construction, and Burisch (1984) for a review and synthesis of research on this topic.

Exercise 2



Decide what the correct level of measurement is for each of the examples below.

	Nominal	Ordinal	Interval	Ratio
Course grades (A, B, C, D)	□	□	□	□
Commute times (minutes)	□	□	□	□
Health status (Excellent, Average, Poor)	□	□	□	□
Gender	□	□	□	□
Age	□	□	□	□
High school test scores (such as 78, 80, or 95 percent correct)	□	□	□	□
Type of surgery	□	□	□	□



	Nominal	Ordinal	Interval	Ratio
Difficulty (Impossible, Difficult, Easy)	□	□	□	□
Hours spent exercising in one week	□	□	□	□
Body temperature in degrees Fahrenheit (F)	□	□	□	□
Race/ethnicity	□	□	□	□
Breast cancer rates in geographic regions (cases per 100,000) people	□	□	□	□
Test scores, such as an IQ score of 90 or 115 on an intelligence test	□	□	□	□

6. Problems in Measuring Constructs

Normality of distributions

Many analytic options (correlation, regression, factor analysis) require the assumption that item scores are (approximately) normally distributed (McDonald, 1999; Nunnally & Bernstein, 1994).

Normal distribution

- Bell-shaped
- Mean and Variance describe distribution well
- Skewness and kurtosis are minimal

Non-normal distribution

Skewed or kurtotic distributions

- **Skewness:** depart from bell-shaped curve by having longer tail at one end of distribution
  - **Positive skew** – long tail at high end of scale
  - **Negative skew** – long tail at low end of scale

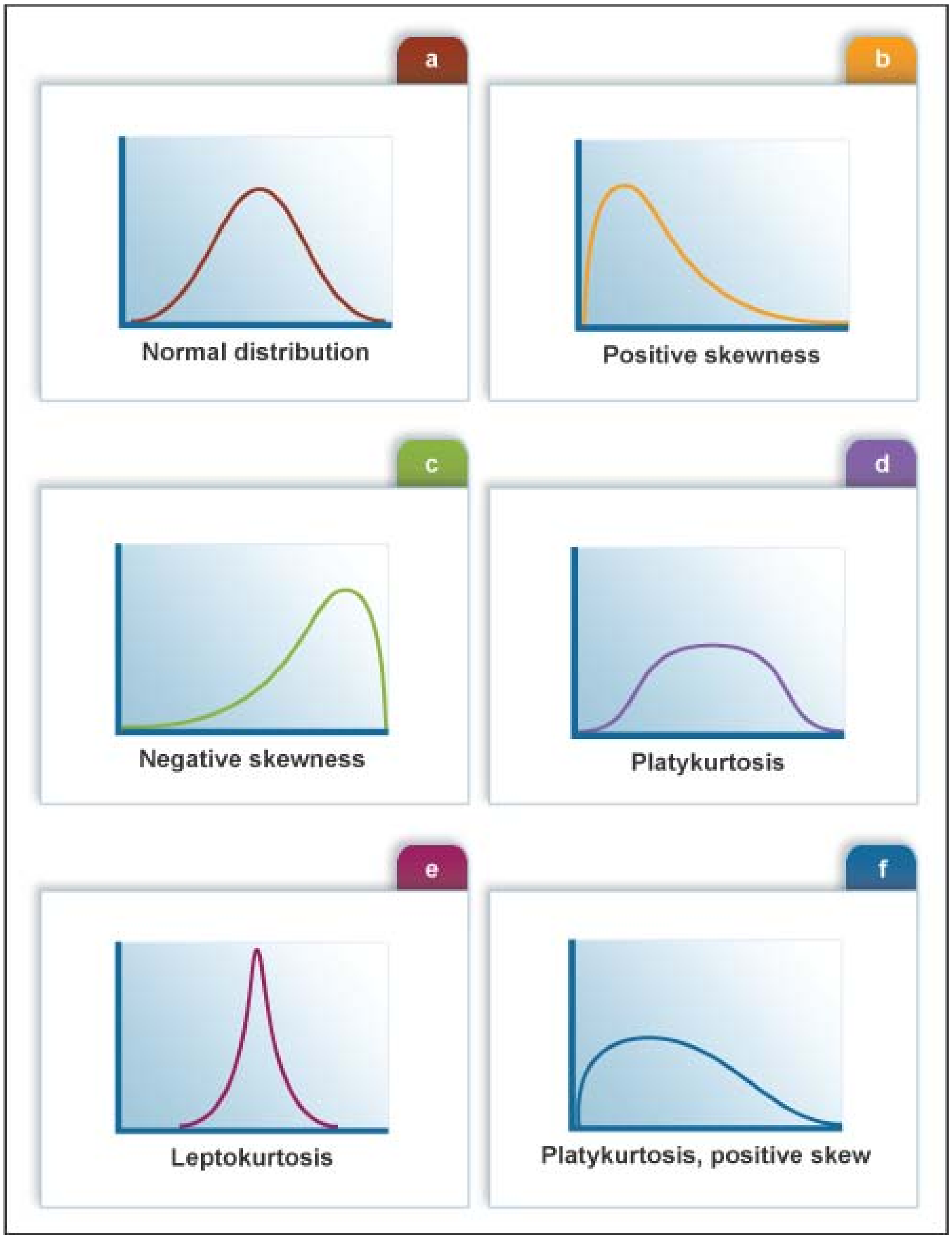
Example 21

Skewness example:

Self-esteem is often negatively skewed because people tend to have positive self evaluations, so few persons use low end of scale.

- **Kurtosis:** depart from bell-shaped curve by having either heavy or light tails (either many or few responses, respectively, at the high and low ends of scale)
  - **Leptokurtosis** – higher, narrower peak of distribution (relative to bell-shaped curve) and therefore “fatter” tails (or more persons scoring in extreme range of values)
  - **Platykurtosis** – lower, wider peak of distribution (relative to bell-shaped curve) and therefore “thinner” tails (or fewer persons having extreme scores)

Figure 1


a
Normal distribution
b
Positive skewness
c
Negative skewness
d
Platykurtosis
e
Leptokurtosis
f
Platykurtosis, positive skew

6. Problems in Measuring Constructs

Dealing with non-normality

Various options can be used to deal with non-normality, including:

Transforming values

If distribution is:

- Negatively skewed, squaring values may make the distribution more normal.
- Positively skewed, then a log transformation or square root transformation may make the distribution more normal.

Revising items to make content more (or less) extreme

If distribution is:

- Negatively skewed, make the item harder to endorse.

Example 22

Negatively skewed example:

Consider the item "*I am usually happy*." assessed on 1-to-7 scale.

- Most persons might use scale points 5-7.
- Few would use scale points 1-4.

Reword the item "*I am the happiest person I know*."

- Fewer persons would use the highest scale points.
- More would use lower scale points.

- Positively skewed, make the item easier to endorse.

Example 23

Positively skewed example:

Consider the item “*I often use harsh language when disciplining my children*.” again assessed on 1-7 scale.

- Few persons would use high scale points (e.g., 4-7).
- Most persons would use rather low scale points.

Reword the item “*I occasionally use harsh language when disciplining my children*.”
More persons are likely to use higher scale values, admitting to using such language only occasionally.

6. Problems in Measuring Constructs

Biases in Measurement

Biases represent systematic influences on item or scale scores that are unrelated to the construct to be measured (Paulhus, 1991). Bias falls under the rubric of construct-irrelevant variance.

Bias in Self-Report

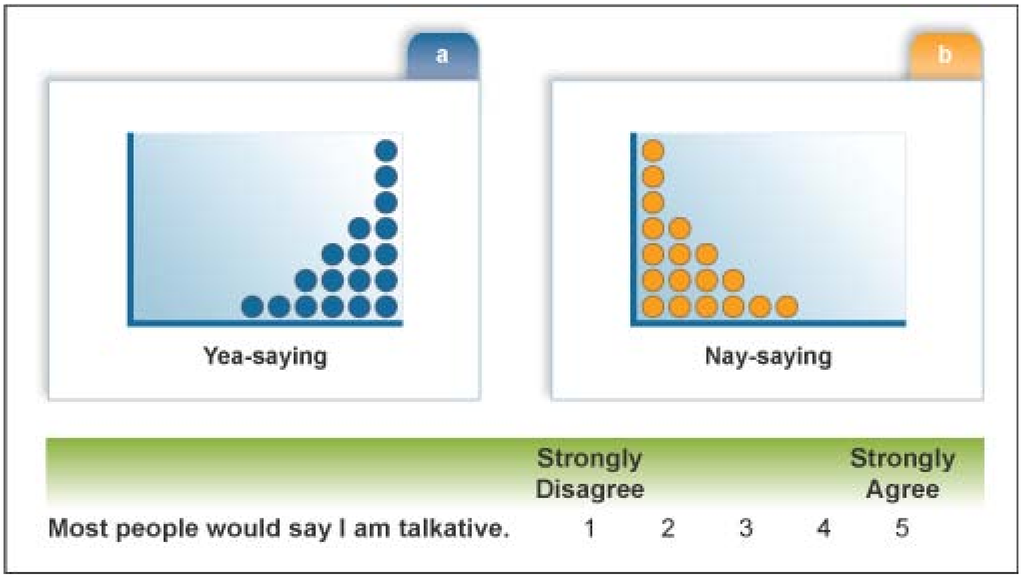
Self-reports are associated with several forms of potential bias:

Acquiescence: “yea-saying” or “nay-saying”.

- “Yea-saying” – tendency to endorse items regardless of item content.
- “Nay-saying” – tendency to refuse to endorse items regardless of item content.

Solution: Balance item content on a scale, with about half of the items positively worded (so a high score on an item indicates high standing on the trait) and about half of the items negatively worded (so a high score on an item indicates a low standing on the trait).

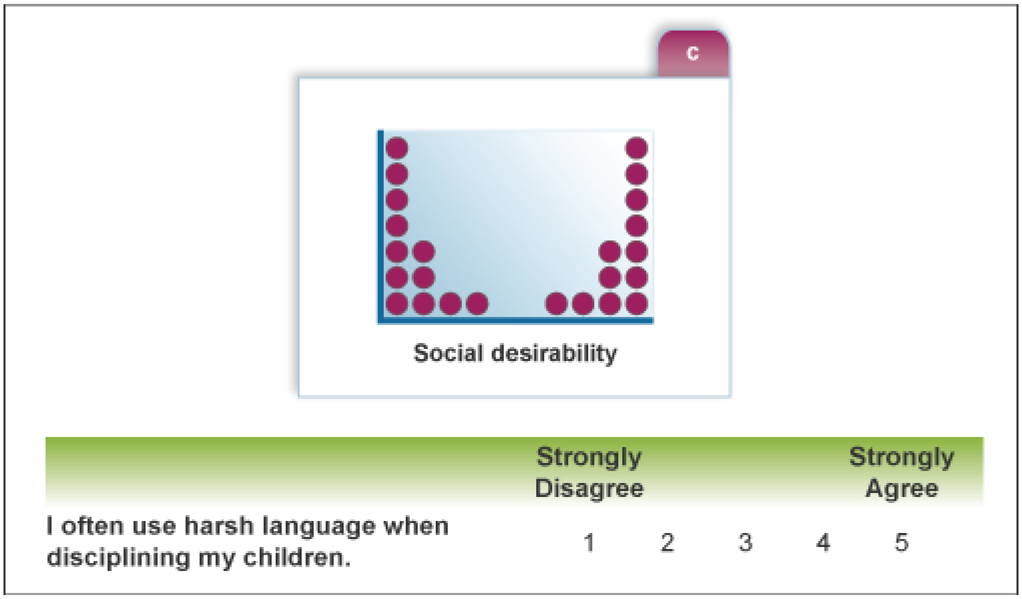
Figure 2



Social desirability: tendency to respond positively to positively valued item content and, conversely, to respond negatively to negatively valued item content (Paulhus, 1984).

> **Solution:** During test development, administer a social desirability scale along with items for scales being developed. Then, during item analyses, one can discard items that are too saturated with social desirability (i.e., that correlate too highly with social desirability).

Figure 3



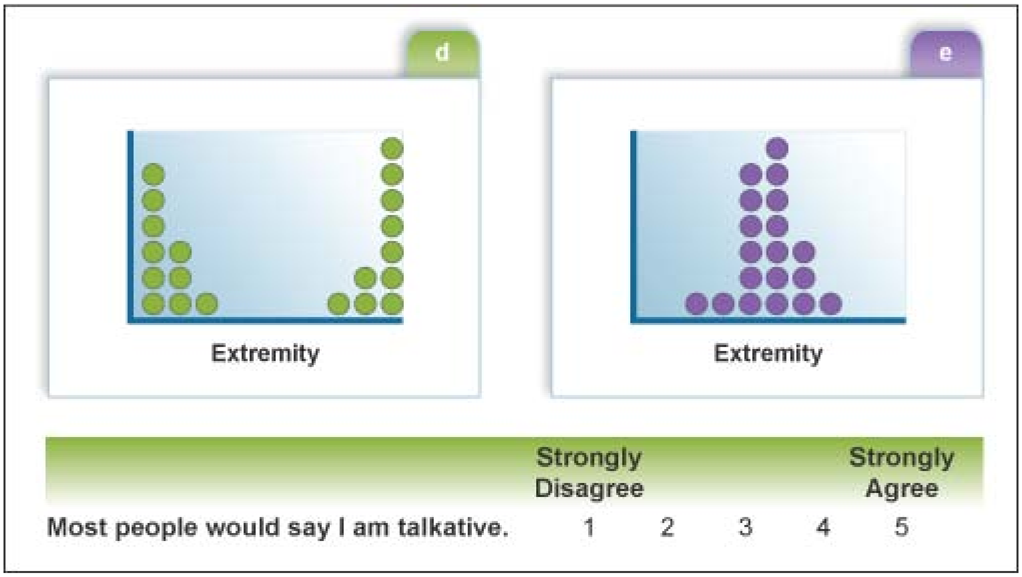
	Strongly Disagree				Strongly Agree
I often use harsh language when disciplining my children.	1	2	3	4	5

Extremity (Peabody, 1962):

- Consider a 1-to-9 scale ranging from “strongly disagree” to “strongly agree”.
- Some persons tend to use only the most extreme values (tend to answer questions using item options 1 or 2 when disagree, or 8 or 9 when agree), so do not use the middle values 3-7 very much.
- Other individuals tend to refrain from using the extreme values even when they very strongly agree or disagree.

Solution: Make sure to have balanced set of items with regard to direction of wording (positively worded and negatively worded), as this will tend to decrease extremity bias effects as well as acquiescence bias effects.

Figure 4



6. Problems in Measuring Constructs

Biases in Measurement

Bias in other-person report

Biases may also arise when other persons (e.g., parents, teachers, observers) provide reports on an individual (Campbell & Fiske, 1959; Eid & Diener, 2006).

- Anchoring: Bias in ratings due to prior information.

If raters are informed that “many people exhibit a given type of behavior” (or tend not to exhibit that behavior), this may influence their judgments of other persons’ actions and behaviors.

Solution: To avoid idiosyncratic anchors, provide explicit anchors for raters.

Example 24

Anchoring example:

An example might be: “In comparison to other persons of the same age and sex, please rate the person on the following scales: …”.

- **Halo effects:** Bias that yields a generalized positive or negative evaluation.

Halo effects arise in many situations in which the rater knows the person being rated rather well.

Solution: Obtain ratings of each target from multiple raters. Thus, one could obtain ratings by multiple supervisors of each target subordinate, or could obtain ratings by both mother and father of target children. Sophisticated methods of analysis can then be used to model the effects of halo bias, but only if two or more raters have been used to obtain ratings.

Example 25

Halo effect examples:

Supervisors may have generalized positive or negative biases regarding their subordinates. Or, parents may have generalized positive or negative biases about one or another of their children.

6. Problems in Measuring Constructs

Base rate issues (Cohen, 1994)

How common is the outcome?

- Some outcomes are rather common
    - Depression
    - High blood pressure
- Other outcomes are less common
    - henylketonuria (PKU)

Uncommon outcomes require careful study to ensure that results from an empirical study generalize to the population.

- Sensitivity - proportion of persons with a given condition who are correctly identified by a test (or scale) as having the condition.
- Specificity - proportion of persons who do NOT have the condition who are correctly identified by a test as NOT having the condition.

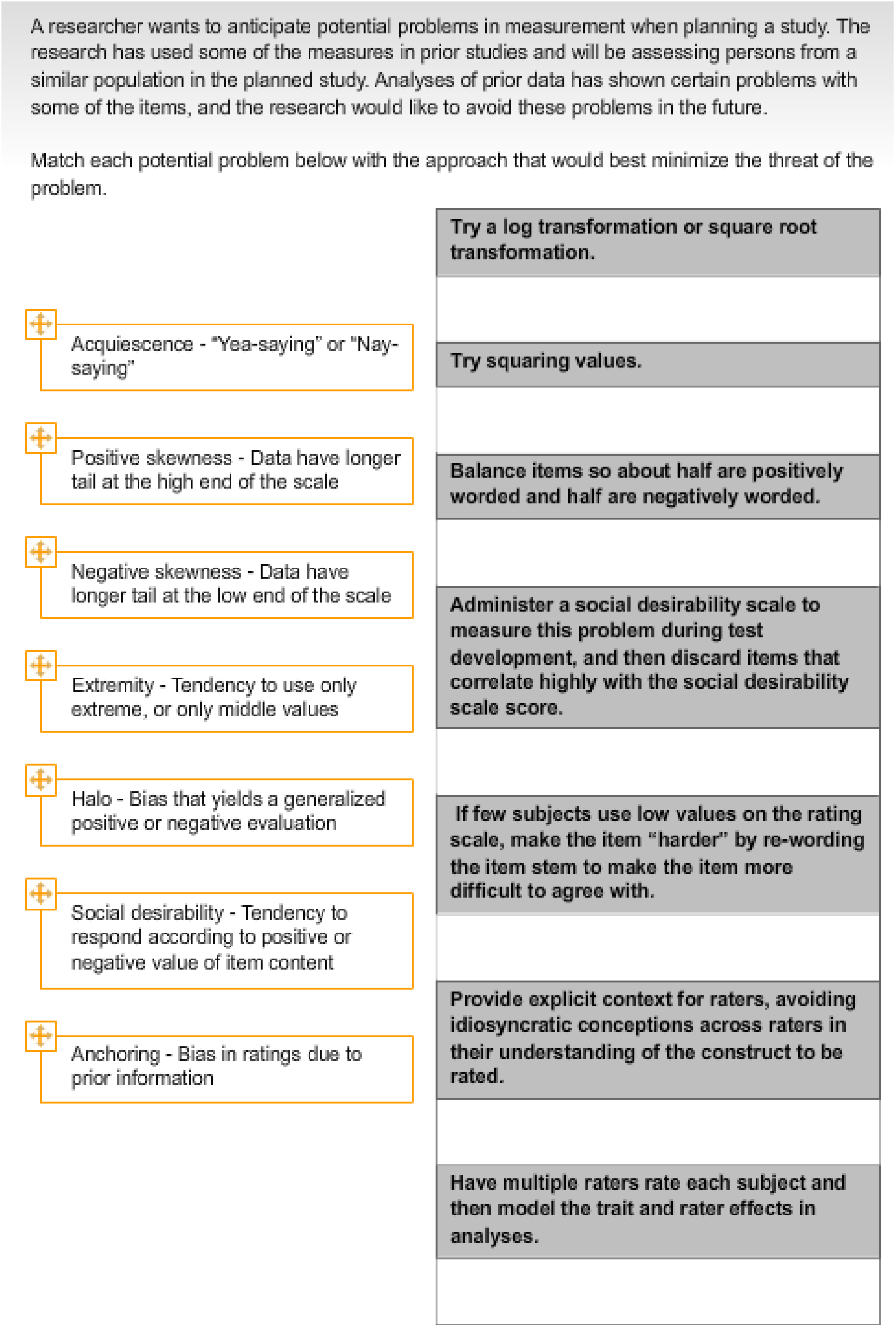
Solution - Use representative sample of persons from population – using a representative sample of participants from the population will ensure that various outcomes, especially sensitivity and specificity of measurement, will reflect likely results in the population, but very large samples may be required to ensure representation of uncommon outcomes.

Or

Extend results from the sample to the population, using base rate information – if an outcome of interest is very unlikely, a representative sample from the population may have to be very large to ensure a sufficient number of index cases to yield stable estimates. In such cases, one might oversample certain strata in the population to make sure that stable estimation is possible. Then, during the analysis phase, one can re-weight cases to obtain appropriate estimates of results that acknowledge the information regarding population base rates.

6. Problems in Measuring Constructs

Exercise 3



A researcher wants to anticipate potential problems in measurement when planning a study. The research has used some of the measures in prior studies and will be assessing persons from a similar population in the planned study. Analyses of prior data has shown certain problems with some of the items, and the research would like to avoid these problems in the future.

Match each potential problem below with the approach that would best minimize the threat of the problem.

- Acquiescence - "Yea-saying" or "Nay-saying"
- Positive skewness - Data have longer tail at the high end of the scale
- Negative skewness - Data have longer tail at the low end of the scale
- Extremity - Tendency to use only extreme, or only middle values
- Halo - Bias that yields a generalized positive or negative evaluation
- Social desirability - Tendency to respond according to positive or negative value of item content
- Anchoring - Bias in ratings due to prior information

Try a log transformation or square root transformation.

Try squaring values.

Balance items so about half are positively worded and half are negatively worded.

Administer a social desirability scale to measure this problem during test development, and then discard items that correlate highly with the social desirability scale score.

If few subjects use low values on the rating scale, make the item "harder" by re-wording the item stem to make the item more difficult to agree with.

Provide explicit context for raters, avoiding idiosyncratic conceptions across raters in their understanding of the construct to be rated.

Have multiple raters rate each subject and then model the trait and rater effects in analyses.

7. Reliability

Reliability refers to the precision with which a scale or instrument assesses a dimension. If one were to administer a scale twice to a sample of participants, one would not expect to obtain precisely the same score for each participant at the two occasions. But, the closer each person's score at time 1 corresponds to his or her score at time 2, the higher the reliability of the measure. Thus, reliability refers to the reproducibility of scores on multiple, theoretical applications of the measuring instrument (McDonald, 1999).

Reliability is defined as the proportion of variance in observed test score that is related to true scores (Cronbach, 1951; McDonald, 1999). Under classical test theory (see above), we can distinguish three sources of variance: (a) true score variance, (b) error variance (or measurement error), and (c) total scale variance (which is the sum of true score and error variance).

Unfortunately, we typically do not have estimates of true score variance or error variance, having only an estimate of the total scale variance. But, invoking simple assumptions regarding the true and error scores (see preceding section of Classical Test Theory), we can obtain estimates of the ratio of true score variance to total score variance, and we use these estimates as our estimates of the reliability of a scale.

TYPES OF RELIABILITY

Parallel Forms Reliability (Or Coefficient of Equivalence)

Correlation between two separate forms or scales developed to assess the same construct.

This coefficient tells us: precision of measurement of each form – how accurately individuals are characterized at a single point in time by each form.

Parallel forms are difficult to develop, must meet exacting standards. They must have:

- Equal means
- Equal standard deviations
- Equal reliabilities

- Equal correlations with outside variables

The correlation between the two parallel forms is the parallel forms reliability for either of the parallel forms, using the other as parallel form. For example, assume the presence of Forms A and B to measure a given construct. If a sample of individuals is administered both forms and scores on the two forms correlate .84, then the parallel forms reliability for Form A is .84 (using Form B as the basis of this claim) and the parallel forms reliability for Form B is .84 (using Form A as the basis of this claim).

Split-Half Reliability

Correlation between two scores constructed as (random) halves of a set of items on a scale (e.g., the correlation of the sum score on even-numbered items with the sum score on odd-numbered items).

This coefficient tells us: precision of measurement of the form in question at the given time of measurement.

Assume the presence of two split halves – Half E (for even-numbered items) and Half O (for odd-numbered items). Assume that the correlation between Half E and Half O is signified as rHO, and rHO = .80. This reliability of .80 is the reliability of each half of the scale. But, because the scale is twice as long as either half (if each half of the scale consists of 6 items, the total scale has 12 items), this reliability must be adjusted upward to estimate the reliability of the total scale, or rxx. To do so, one must use a special form of the Spearman-Brown prophecy formula.

$$r_{xx(2)} = \frac{2\,r_{HO}}{1+r_{HO}} = \frac{2(.80)}{1+.80} = \frac{1.60}{1.80} = .89$$

In the preceding formula, the term rxx(2) has a (2) in its subscript to indicate that it is an estimated reliability assuming that the total scale is 2 times as long as either form that was correlated.

Internal Consistency Reliability

Touted as average of all possible split-halves, based on the set of items responses, specifically the variance-covariance matrix among items.

This coefficient tells us: precision of measurement of the form in question at the given time of measurement.

- Kuder-Richardson 20, 21 for dichotomous items
- Coefficient alpha
- Coefficient omega

Test-Retest Reliability (Or Coefficient of Stability)

Correlation between scores for the same individuals at two different points in time.

This coefficient tells us: whether individual differences are similar at the two points in time.

- **Trait construct** – ideally, would have high stability
- **State construct** – ideally, would have low(er) stability

Interval is important.

- **Interval too short** – too much “memory” effect – person remembers how s/he answered last time, answers the same the next time
- **Interval too long** – construct (true score) has changed, and change in construct would be treated as error variance

Coefficient of Stability and Equivalence

Correlation between scores for same individuals at two points in time, in which individuals take one parallel form at one time and the different parallel form at the second time.

This coefficient tells us: whether individual differences are similar at the two points in time even when precise content of measure changes.

Interrater Reliability

Have 2 (or more) raters provide ratings of each person in a sample (e.g., ratings of aggressiveness of each child based on observing the children interacting on a playground for several days). Then correlate ratings by rater 1 with ratings by rater 2.

This coefficient tells us: Whether two (or more) individuals have similar views of differences among persons being rated (i.e., whether they “see” the same things in people).

Problem: Interrater reliability may be high, even though certain other forms of reliability are low.

7. Reliability

Standards for reliability

Standards for acceptable levels of reliability vary across experts, but some general guidelines can be provided (Clark & Watson, 1995; Nunnally & Bernstein, 1994):

Table 1

General Reliability Standards	
.90 or higher	Excellent
.80 to .90	Strong
.70 to .80	Acceptable
.60 to .70	Weak or Poor
below .60	Unacceptable

The preceding standards for evaluating reliability are reasonable for multi-item scales that contain 8-10 items or more, particularly if the rating for each item is on a Likert scale with a 1-4 or 1-5 or larger scale.

Scales with few items or with more restricted rating scales (e.g., dichotomous scales) will likely have fewer points of discrimination among participants and this may lead to lower levels of reliability (e.g., reliability between .50 and .60). Such scales may still be usable for research purposes, but the “proof is in the pudding” (i.e., usability of such scales will depend on whether they are strongly related to other measures as hypothesized.

In addition, measuring devices to be used in high-stakes decision-making, such as deciding whether a person has mental retardation, should have very high levels of reliability, preferably above .95. Traditional individually administered tests of intelligence tend to attain this level of reliability.

8. Validity

Validity involves, in its broadest construal, whether a scale assesses the construct it was intended to assess (Cronbach & Meehl, 1955).

Traditional Tripartite Delineation of Validity – The 3C's of Validity

Content Validity:

Judgment by experts that domain of interest has been properly sampled

Criterion-Related Validity:

Correlation of test score with one or more criteria it should predict:

- **Predictive Validity** – test score correlates with criterion obtained at a later point in time

Example 27

Concurrent validity example:

Score on Stanford Binet test of intelligence correlates with score on the Wechsler Adult Intelligence Scale.

Construct Validity:

Validity of the use of a score for an intended purpose – the core of the validation of a measure.

For example, does it:

- Correlate with other measures of the same purported construct?
- Correlate with criteria it is supposed to correlate with?
- Vary appropriately across contexts (e.g., anxiety scores increase when confronted with anxiety provoking situation)?
- Yield appropriate factor analytic solution? If one hypothesizes a single dimension is being assessed, does a factor analysis show that a one-factor model is optimal?

8. Validity

Convergent and discriminant validity (Campbell & Fiske, 1959)

Convergent validity

Does a measure correlate highly with other purported measures of the same construct?

Discriminant validity

Does a measure correlate at lower levels with measures of different constructs?

Formalized in the multitrait-multimethod matrix (MTMM matrix) (Campbell & Fiske, 1959)

- Initially, informal methods were used to analyze patterns in MTMM matrices (circa 1959 – 1970).
- Now, sophisticated factor analytic models can be use to represent patterns in MTMM matrices.
- Can estimate proportion of variance in a measure that is trait-related (or construct-related) and proportion that is method-related (or construct irrelevant, but reliable variance)
- Can evaluate different types of method effects, specifically effects associated with:
  - Scale construction methods
  - Raters (child self-report, mother report, father report, teacher report)

9. Summary

Measuring subjective phenomena is a task fraught with potential problems, but many researchers must engage in such measurement to meet a host of theoretical and practical goals. On the theoretical side, testing theories in psychology and related behavioral sciences requires us to measure personal, covert, subjective constructs. In practical applications, researchers in many different fields must have measures of subjective phenomena to answer their research questions. For example, in medical research, investigators may wish to measure pain that a patient is experiencing right now or over the past week or month, and may wish to develop a measure that is optimally responsive to treatment effectiveness. Or, researchers of cognitive and brain-related changes during aging may need measures of loneliness and perceived social support to investigate the moderating effects these constructs have on the rate of age-related change.

Although measuring subjective phenomena has attendant problems, careful consideration of the goals of measurement and of the various potential biases in measurement should ease the burden of the investigator or research team in the development of new measures. Accurate measurement begins with a careful definition of the construct to be measured and the ways in which the construct can be exemplified in behavior, emotional responses, or other ways. Items should be developed that map onto the varied phenomena associated with the construct. Potential sources of bias should be investigated. Over a half century of research in psychology has documented sources of bias in self-reports, other-reports, and observer ratings. These potential biases can be exploited in research studies to study whether scale items are systematically biased in any way; follow-up studies can be conducted to counter these biases. Once a scale has been developed, assessments of psychometric properties, such as reliability, should be conducted, and the applicability of the scale to different populations should be investigated.

The present chapter provides an introduction to the process of measuring subjective phenomena. But, the development of good, high-quality measures of subjective phenomena cannot be learned only from books, but must be learned through the process of the development and critical examination of measures. This chapter offers ways to think about examining measures for their strong and weak points and should be useful to researchers who intend to develop new measures or improve existing measures of subjective phenomena.

10. References

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole.

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. American Psychologist, 39, 214-227.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull, 56(2), 81-105.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7, 309-319.

Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49, 997-1003.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychol Bull, 52(4), 281-302.

DeVellis, R. F. (2003). Scale development: Theory and applications (2nd ed.). Thousand Oaks, CA: Sage.

Donnellan, M. B., Lucas, R. E., & Fleeson, W. (2009). Introduction to personality and assessment at age 40: Reflections on the legacy of the person-situation debate and the future of person-situation integration. Journal of Research in Personality, 43, 117-119.

Eid, M., & Diener, E. (Eds.). (2006). Handbook of multimethod measurement in psychology. Washington, DC: American Psychological Association.

Gaudry, E., Vagg, P., & Spielberger, C. D. (1975). Validation of the state-trait distinction in anxiety research. Multivariate Behavioral Research, 10, 331-341.

Hampson, S. E., & Goldberg, L. R. (2006). A first large cohort study of personality trait stability over the 40 years between elementary school and midlife. J Pers Soc Psychol, 91(4), 763-779.

Hase, H. D., & Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventory scales. Psychol Bull, 67(4), 231-248.

Horn, J. L., & Hofer, S. M. (1992). Major abilities and development in the adult period. In R. J. Sternberg & C. A. Berg (Eds.), Intellectual development (pp. 44-99). New York: Cambridge University Press.

Jackson, D. N. (1971). dynamics of structured personality tests: 1971. Psychological Review, 78, 229-248.

Lucas, R. E., & Donnellan, M. B. (2009). If the person-situation debate is really over, why does it still generate so much negative affect? Journal of Research in Personality, 43, 146-149.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Erlbaum.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741-749.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill.

Paulhus, D. L. (1984). Two-component models of social desirable responding. Journal of Personality and Social Psychology, 46, 598-609.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), Measures of personality and social psychological attitudes (pp. 17-59). New York: Academic.

Peabody, D. (1962). Two components in bipolar scales: direction and extremeness. Psychol Rev, 69, 65-73.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. J Pers Soc Psychol, 94(4), 718-737.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? Journal of Personality and Social Psychology, 84, 1041-1053.

Waller, N. G., & Meehl, P. E. (1998). Multivariate taxometric procedures: Distinguishing types from continua. Thousand Oaks, CA: Sage.

Widiger, T. A., & Trull, T. J. (2007). Plate tectonics in the classification of personality disorder: shifting to a dimensional model. Am Psychol, 62(2), 71-83.

11. Author Biography

Keith F. Widaman, Ph.D., is a Distinguished Professor in the Department of Psychology at the University of California at Davis. A faculty member at UC Davis since 1999, he is also the immediate past Chair of the department. Widaman previously was a member of the faculty for 19 years at the University of California at Riverside (1980-1999). He received his Ph.D. in 1982 from the Ohio State University, with major emphasis in Developmental Psychology and a minor in Quantitative Psychology. He has extensive experience in the use of multivariate linear models, including regression analysis, factor analysis, structural equation modeling, and the modeling of longitudinal data. His substantive program of research focuses on family, economic, cultural, and other influences on child development and the structure and development of mental abilities and everyday skills and abilities in both representative and developmentally disabled populations. He has published extensively in methods-oriented journals such as Psychological Methods, Multivariate Behavioral Research, and Applied Psychological Measurement, and in substantive journals such as Developmental Psychology, the American Journal on Mental Retardation, Child Development, and Intelligence.

Widaman has served on the Editorial Boards of many journals, including Psychological Methods, Multivariate Behavioral Research, the Journal of Abnormal Psychology, Psychological Assessment, Intelligence, and Structural Equation Modeling. He is a Fellow of the American Psychological Association (Divisions 5, 7, and 33) and the Association for Psychological Science. He received the 1992 Raymond B. Cattell Award for early career contributions to multivariate psychology from the Society of Multivariate Experimental Psychology (SMEP), has twice received the Tanaka Award for best article in SMEP's journal Multivariate Behavioral Research, and is a Past President of the society. He was a member of the National Research Council Committee on Disability Determination in Mental Retardation (2000-2002), which produced a monograph advising the US Social Security Administration on structuring diagnostic practices regarding mental retardation, and he wrote the chapter on adaptive behavior in the American Psychological Association Manual of Diagnosis and Professional Practice in Mental Retardation. Widaman is a member of a working group that is developing a new Diagnostic Adaptive Behavior Scale under the aegis of the American Association on Intellectual and Developmental Disabilities. He is also an advisor to the Developmental Disabilities Work Group responsible for developing diagnostic guidelines for the Diagnostic and Statistical Manual – V (DSM-V) of the American Psychiatric Association. During the late 1990's, he was Chair of the University of

California systemwide Board of Admissions and Relations with Schools, which sets standards for eligibility for admission to the University of California system.