INSTITUTE FOR DEFENSE ANALYSES

Foundations of Psychological Measurement

Heather Wojton

February 2017

Approved for public release.

IDA NS D-8273

Log: H 2016-001298

INSTITUTE FOR DEFENSE
ANALYSES 4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About this Publication

Psychological measurement is an important issue throughout the Department of Defense (DoD). For instance, the DoD engages in psychological measurement to place military personnel into specialties, evaluate the mental health of military personnel, evaluate the quality of human-systems interactions, and identify factors that affect crime rates on bases. Given its broad use, researchers and decision-makers need to understand the basics of psychological measurement – most notably, the development of surveys. This briefing discusses 1) the goals and challenges of psychological measurement, 2) basic measurement concepts and how they apply to psychological measurement, 3) basics for developing scales to measure psychological attributes, and 4) methods for ensuring that scales are reliable and valid.

For More Information:

Laura Freeman, Project Leader
lfreeman@ida.org, (703) 845-2084

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org, (703) 845-2462

INSTITUTE FOR DEFENSE ANALYSES

IDA NS D-8273

Foundations of Psychological Measurement

Heather Wojton

This page intentionally left blank.

Executive Summary

The Institute for Defense Analyses (IDA) periodically receives requests for analysts to develop courses on survey methodology. The attached slides are designed to accompany a 2-hour course on the foundations of psychological measurement and scale development for analysts at IDA and throughout the Department of Defense (DoD). The topics covered in each section of this course are outlined in greater detail below.

A. The Goals & Challenges of Psychological Measurement

The DoD uses psychological measurement to aid in decision-making about a variety of issues, including the mental health of military personnel before and after combat, and the quality of human-systems interactions. To develop quality survey instruments and interpret the data obtained from these instruments appropriately, analysts and decision-makers must understand the factors that affect the reliability and validity of psychological measurement.

Psychologists use surveys to measure human behavior, then make inferences from these behaviors regarding their underlying psychological attributes. In general, the goal of such measurement is to compare the behavior of two or more people at a single point in time, compare the behavior of the same people at different points in time, compare the behavior of people under different conditions, or use some combination of these. To achieve these goals, analysts must develop survey instruments that account for the complex, multi-dimensional nature of psychological attributes and are robust to participant reactivity, experimenter bias, and score sensitivity.

B. Basics of Psychological Measurement

Measurement is the assignment of numerals to objects or events, and numerals can represent psychological attributes in different ways. Numerals that possess the property of identity, for instance, serve strictly as labels of categories, reflecting differences in kind rather than amount. For example, in a group of men and a group of women that are labeled 1 and 2 (respectively), the numerals simply indicate that an individual

is a man or a woman. Numerals that possess the property of order, however, convey information about the relative amount of a psychological attribute that people possess. For example, an instructor pilot may use numerals to rank his or her students in terms of their skill. The numeral 1 would represent the student with the most skill, the numeral 2 would represent the student with the second most skill, the numeral 3 would represent the student with the third most skill, and so on. These numerals tell us which students have more skill and which have less, but they do not tell us how much more or how much less skill they possess than their peers. Numerals with the property of quantity are the only numerals that reflect the actual amount of a psychological attribute that people possess. They possess all the properties of real numbers, and therefore adhere to the rules of addidivity and counting: the number 1 defines the basic unit on the scale, each numeral represents a count of basic units, and all basic units are identical in size.

C. Scale Development Basics

The primary goals of scale development are to construct a scale with measurement units that are clearly defined, closely match the "true" psychological units for the attribute of interest, possess the property of quantity, and do not validate the assumptions of additivity and counting.

Surveys are a set of questions that comprise two parts: item and response-option. The item is the statement that the person responds to, and the response-option is the scale the person uses to respond to the item. Surveys can comprise one or more scales, each of which is designed to measure a single psychological attribute. The psychological attribute may itself be made up of more specific attributes, which are referred to as dimensions or subscales. Typically, scales consist of a series of questions that are combined (added or summed) in some way to create a total score. These total scores are referred to as composites and have several advantages over single-item scales. First, they better represent the complex, multifacted nature of most psychological attributes. Second, scales with more questions tend to yield more reliable estimates. Third, composite scores clearly possess the property of quantity; single-item measures may or may not.

Scales can contain one or more subscales (dimensions), and each subscale should reflect a single psychological attribute. For instance, the NASA-Task Load Index (NASA-TLX) is a measure of workload that consists of six subscales that measure more specific aspects of workload, such as mental, physical, and temporal demand and effort; frustration; and perceived performance. When developing a scale, analysts should determine how many subscales are in the larger scale. If there are more than one subscale, analysts should also determine whether the subscales are correlated and what each of the subscales means, because the answers to these questions have important implications for how the scale should be scored

and evaluated. Factor analysis, a statistical procedure, is helpful in answering these questions.

Scales that comprise a single subscale are referred to as *unidimensional scales*. A single composite score is computed to reflect the single psychological attribute it measures, and the quality (reliability and validity) of the scale (not the individual questions) is evaluated for the composite score. A conceptual representation of a unidimensional scale is presented in Figure 1. It is important to note that the psychological attribute is believed to cause responses to the items that comprise the subscale.

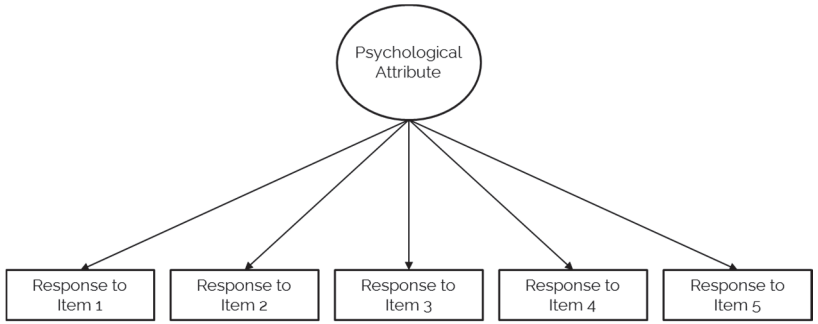


Figure 1. Unidimensional Scale

Scales with multiple correlated subscales are referred to as *scales with higher-order factors*. Each subscale assesses a different construct that reflects different aspects of a broader psychological attribute, and each subscale is itself unidimensional. Because the subscales are correlated, these scales produce various scores, including a score for each subscale and a total score combined across subscales. Scale quality is evaluated for each composite score. A conceptual representation of a scale with higher-order factors is presented in Figure 2.

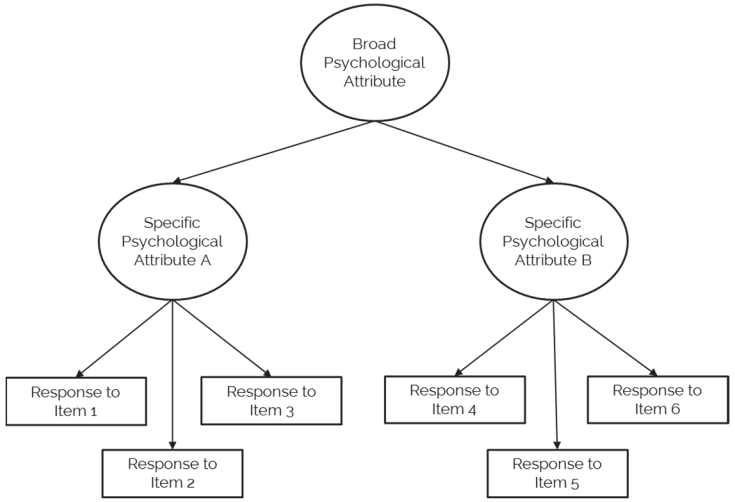


Figure 2. Scale with Higher-Order Factors

Multidimensional scales with uncorrelated dimensions are similar to scales with higher-order factors except that the subscales are uncorrelated. That is, they are not linked by a broader psychological attribute. Practically speaking, the subscales are a set of unrelated unidimensional scales whose items are mixed together. Consequently, composite scores are computed for each subscale, but no total score is computed. The quality of the scale is determined by the reliability and

validity of each subscale score. A conceptual representation of a multidimensional scale with uncorrelated dimensions is presented in Figure 3.

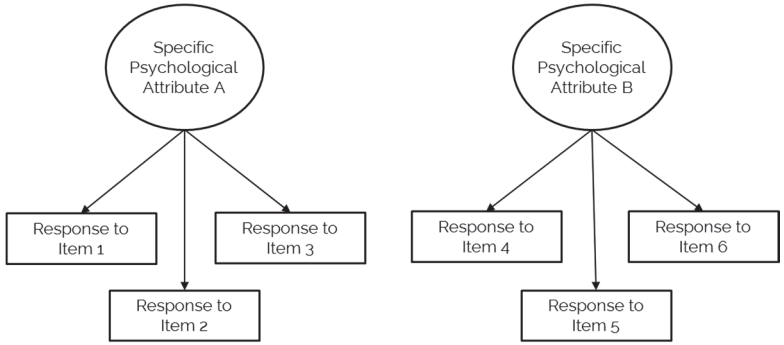


Figure 3. Multidimensional Scale with Uncorrelated Dimensions

D. Reliability & Validity Testing

The quality of a survey is determined by the extent to which it is reliable and valid. Reliability is the extent to which scale scores are a function of respondents' true psychological differences as opposed to measurement error. Validity is the extent to which scale scores measure what they are intended to measure.

There are three primary methods for evaluating reliability: alternate forms reliability, test-retest reliability, and internal consistency reliability. The method chosen for reliability testing depends on the type of data available to the analyst. Likewise, validity can be assessed in a variety of ways, the most rigorous of which include assessing convergent and predictive validity.

Foundations of Psychological Measurement

How to develop valid and reliable scales

Heather Wojton, Ph.D.

Institute for Defense Analyses

10/26/2016

This page intentionally left blank.

The Mind Reader's Toolbox

Surveys are a form of psychological measurement

(except purely demographic surveys)

Nearly **everyone** in industrialized countries is affected by psychological measurement at some point in their lives.

- Standardized knowledge and intelligence tests in education
- Personality tests in the hiring process
- Political polls
- Death penalty

The Department of Defense engages in psychological measurement to:

- Place military personnel into specialties
- Evaluate the mental health of military personnel
- Evaluate the quality of human-system interactions
- Identify factors that affect crime rates on military bases

Researchers **must** understand the properties that affect psychological measurement to develop quality surveys

Course Objectives

1. Identify psychological measurement's goals and challenges
2. Understand basic measurement concepts and how they apply to psychological measurement
3. Understand scale development basics
4. Understand the importance of reliability and validity testing scales, factors that affect reliability and validity, and how to conduct reliability and validity testing

This page intentionally left blank.

Measuring Properties of the Physical World

Scientists use different kinds of instruments to measure **observable** properties of the physical world

For example, we use a tape measure to measure an object's height

Scientists also use instruments to measure **unobservable** properties of the physical world

For example, we use clocks to measure time

Psychological Measurement is the Same

Psychologists use instruments to measure observable properties of the physical world – namely, **behavior**.

Instruments: surveys, psychological tests, physiological measures (e.g., blood pressure cuff), counters (etc.)

Because they are interested in the behavior…

For example, facial expressions or sleep patterns

OR to assess unobservable psychological attributes

For example, intelligence, workload, stress, or memory

Connecting the Observable to the Unobservable

Simplified method for assessing psychological attributes:

1. Identify a behavior believed to represents a specific psychological attribute, state, or process
2. Measure the behavior
3. Interpret the measurement in terms of the underlying psychological attribute, state, or process

Surveys are often developed to sample the behavior believed to be sensitive to the underlying psychological attribute

What behavior do surveys sample?

Working Memory Example

How do you measure working memory?

For example, we might flash a wordlist on the screen for 5 seconds. Then, ask 2 groups of participants to write down as many words as they can remember

If performance differs, then we might assume that the groups differ in their working memory capacity

link	rule	horizon
win	slim	timetable
add	opportunity	elephant
cup	platinum	cathedral
list	livelihood	computer
knot	overestimate	mouse
spade	regiment	pencil
watch	government	elevator

Notice: we made an inference from an observable behavior to an unobservable psychological attribute.

To be valid, the behavior must be **theoretically linked** to the psychological attribute.

A Note on Systematic Sampling

Behavior must be sampled systematically to be useful

Typically, samples of behavior are collected to:

1. Compare the behavior of 2 or more people at the same point in time
2. Compare the behavior of the same people at different points in time
3. Compare the behavior of people under different conditions

These goals may be explicit or implicit

Psychometrics

The science that evaluates the attributes of surveys and other types of psychological measurement

Psychometrics is concerned with 3 types of information

1. The type of data generated by the measurement instrument
   - Surveys, for example, generate scores or ratings
2. The **reliability** of the data
3. The **validity** of the data

Challenges in Psychological Measurement

Psychological phenomenon are complex

- Use of composite scores

Participant reactivity

- Demand characteristics
- Social desirability
- Malingering

Score sensitivity

☐ No

☐ Yes

vs.

Strongly Disagree				Strongly Agree
1	2	3	4	5

This page intentionally left blank.

Course Objectives

1. Identify psychological measurement's goals and challenges
2. Understand basic measurement concepts and how they apply to psychological measurement
3. Understand scale development basics
4. Understand the importance of reliability and validity testing scales, factors that affect reliability and validity, and how to conduct reliability and validity testing

This page intentionally left blank.

"Measurement is the assignment of numerals to objects or events according to rules."

(Stevens, 1946)

The "events" of interest in psychological measurement are people's behaviors

Basic Measurement Concepts

Scaling is the process by which numbers are assigned to represent the quantities of psychological attributes

To appreciate the concept of scaling you must understand:

1. The meaning of numerals
2. How numerals can be used to represent psychological attributes
3. Problems associated with trying to connect numerals and psychological attributes

Strongly Disagree	Somewhat Disagree	Neither Agree Nor Disagree	Somewhat Agree	Strongly Agree
1	2	3	4	5

Numerical Properties

The properties of *identity*, *order*, and *quantity* reflect key differences in how numbers represent psychological attributes

Identity is the most fundamental form of measurement reflecting "sameness" vs. "differentness"

For example, you might ask a teacher to identify children in their class that have behavior problems.

Children that are identified as having behavior problems should be **similar to** each other with respect to their behavior

Children that are identified as having behavior problems should be **different from** those who are classified as not having behavior problems

Property of Identity

Requires that people be sorted into at least 2 categories

Rules for sorting people into categories:

1. People within a category must be identical with respect to the factor used for classification
2. Categories must be mutually exclusive
3. Categories must be exhaustive

The numerals serve strictly as labels of categories, **reflecting differences in kind** rather than amount

Our example doesn't represent the amount of behavior problems, but the presence or absence of problems

Property of Order

Conveys information about the **relative** amount of an attribute that people possess

Indicates the rank order of people relative to each other along some dimension

> For example, you might ask an instructor pilot to rank his or her students according to their level of skill
>
> > The instructor pilot might assign the numeral 1 to the student with the most skill, 2 to the student whose skill is superior to all the other students except the first student (etc.)

Numerals that indicate order are (again) essentially labels

Numerals with the property of order tell us more than those with the property of identity, but are still limited

Property of Quantity

Conveys the greatest information

Numerals act as real numbers, reflect the actual amount of an attribute people possess

- The number 1 defines the size of a basic unit on the scale
- Each numeral represents a count of basic units

For example, a pilot that detects a target in 1 second isn't simply faster at detecting targets than a pilot who detects a target in 3 seconds; he or she is precisely 2 units (seconds) faster

Units of measurement are standardized quantities

Real numbers are continuous

Researchers often assume that scores from surveys have the property of quantity

Units of Measurement

The property of quantity requires that units of measurement be clearly defined

Quantitative measurement hinges on our ability to count these units

- In physical measurement, these units are readily apparent

  For example, you might measure the length of a piece of wood with a tape marked off in inches or centimeters

- In psychological measurement, the units are often less obvious

  In surveys, for example, the measurement units are responses to a series of questions

How do we know to what extent responses are related to the psychological attributes themselves?

Measurement Unit Example

Imagine that you want to measure the height of a bookshelf, but you cannot find a tape measure, yardstick, or ruler

How do you measure it?

Create your own system of measurement!

Take Note:

— The size of your measurement unit is arbitrary

— Your measurement unit could be applied to others objects

Additivity & Counting

Counting is central to all attempts at measurement

One fundamental assumption of counting is that the unit size doesn't change

A unit increase at one point in the measurement process must be the same as a unit increase at another point.

Ideally, measurement units on surveys will correspond closely with "true" psychological units.

Care must be taken to ensure that each measurement unit on a scale reflects the same amount of psychological units

Example: Thinking in "usability units"

Levels of Measurement

The properties of numbers are closely related to the levels of measurement proposed by Stevens (1946)

Stevens's levels of measurement are a set of rules that link the properties of numbers to particular types of observations

Levels of Measurement

Property of Numbers	Nominal	Ordinal	Interval	Ratio
Identity	X	X	X	X
Order		X	X	X
Quantity			X	X
Rational Zero				X
Example	*Sex*	*Education*	*Memory*	*Behavior*

Practical Implications

Researchers treat survey data as possessing the property of quantity

- Particularly, aggregated scores obtained from multi-item, Likert-type scales
- This assumption may not be valid, particularly for brief single-item scales

Parametric tests are most appropriate for data possessing the property of quantity

The goal of all scaling procedures is to represent differences between people

Must keep the property of numbers and assumptions of additivity and counting in mind when constructing and interpreting surveys

Course Objectives

1. Identify psychological measurement's goals and challenges
2. Understand basic measurement concepts and how they apply to psychological measurement
3. Understand scale development basics
4. Understand the importance of reliability and validity testing scales, factors that affect reliability and validity, and how to conduct reliability and validity testing

This page intentionally left blank.

Scale Development Goals

The primary goals of scale development are to:

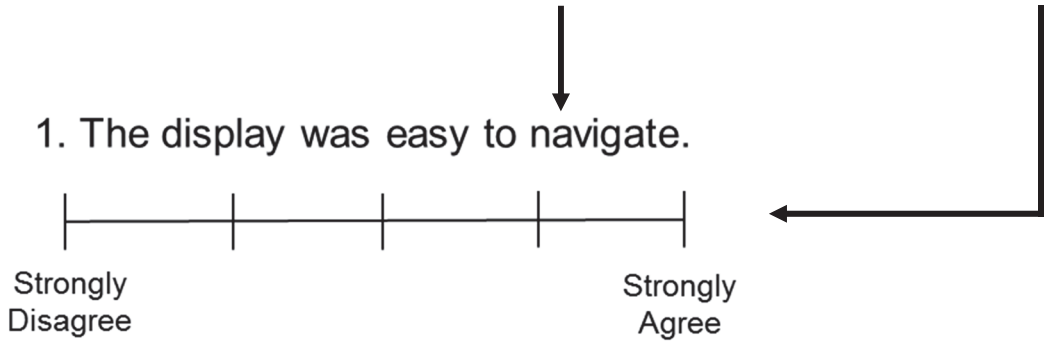
1. Construct a scale with measurement units that are clearly defined and represent the appropriate psychological attribute
2. Construct a scale whose measurement units closely match the “true” psychological units for the attribute of interest
3. Construct a scale that possesses the property of quantity
4. Construct a scale that does not violate the assumptions of additivity and counting

The Survey Instrument

Surveys measure people's attitudes, opinions, feelings

— Comprised of a series of questions

Questions consist of 2 parts, the item and response option



A scale is a set of questions designed to measure the same psychological attribute (thought or feeling)

— Multi-item scales are preferable to single-item scales due to the complexity of psychological attributes

Composite Scores

Typically, responses to items from multi-item scales are combined in some way to create a total score

- Can be summed or averaged

There are three primary advantages to composite scores

1. Better representation of the psychological attribute's complexity
2. Estimates are more reliable (more about that later)
3. Composite scores clearly possess the property of quantity

Dimensionality Example

Imagine that soldiers are asked to take a survey that includes the following six personality traits:

1. Talkative
2. Assertive
3. Imaginative
4. Creative
5. Outgoing
6. Intellectual

They must consider how well each trait describes them

Consider the survey. What does it measure?

— Does it measure 6 separate dimensions of personality or does it measure a single dimension?

Group these traits into clusters based upon similarity

How many clusters or “dimensions” did you create?

Cluster 1

Talkative

Assertive

Outgoing

“Extraversion”

Cluster 2

Imaginative

Creative

Intellectual

“Openness to Experience”

How many clusters or “dimensions” did you create?

Cluster 1	**Cluster 2**	**Cluster 3**
Talkative	Imaginative	Intellectual
Assertive	Creative	
Outgoing		

Dimensionality is a fundamental question in scale development, evaluation, and use.

Scale Dimensionality

In general, when we measure an attribute of a person or object, we intend to measure a single attribute

- For example, you would not add a person's height and hair length together to form a "total" score
- Similarly, adding measures of personality and memory together produces a total score that is meaningless

For this reason, composite scores should reflect a single psychological attribute

- However, a scale may include more than one dimension

  For example, you might have a workload scale that measures 3 dimensions of workload including mental, physical, and temporal.

Scale Dimensionality

Researchers should ask themselves 3 questions about scale dimensionality:

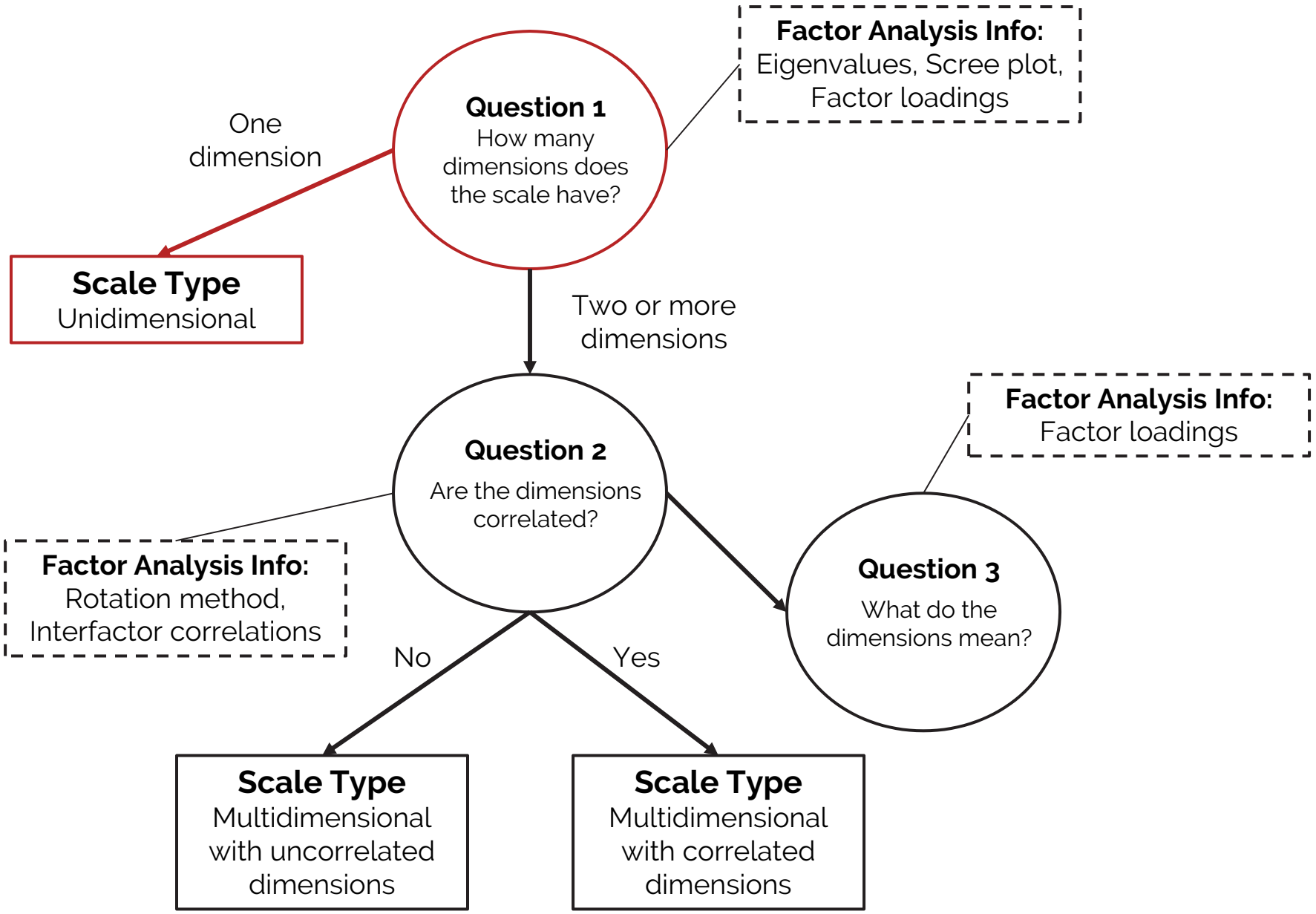
1. How many dimensions are reflected in the scale?
2. If the scale has multiple dimensions, are they correlated with each other?
3. If the scale has multiple dimensions, what are they?

Answers to these questions determine how scales are scored and interpreted

- Including if it's appropriate to compute a "total" scale score for multi-dimensional scales

Factor analysis is useful for determining the number of dimensions in a scale and how they are correlated

Types of Scales

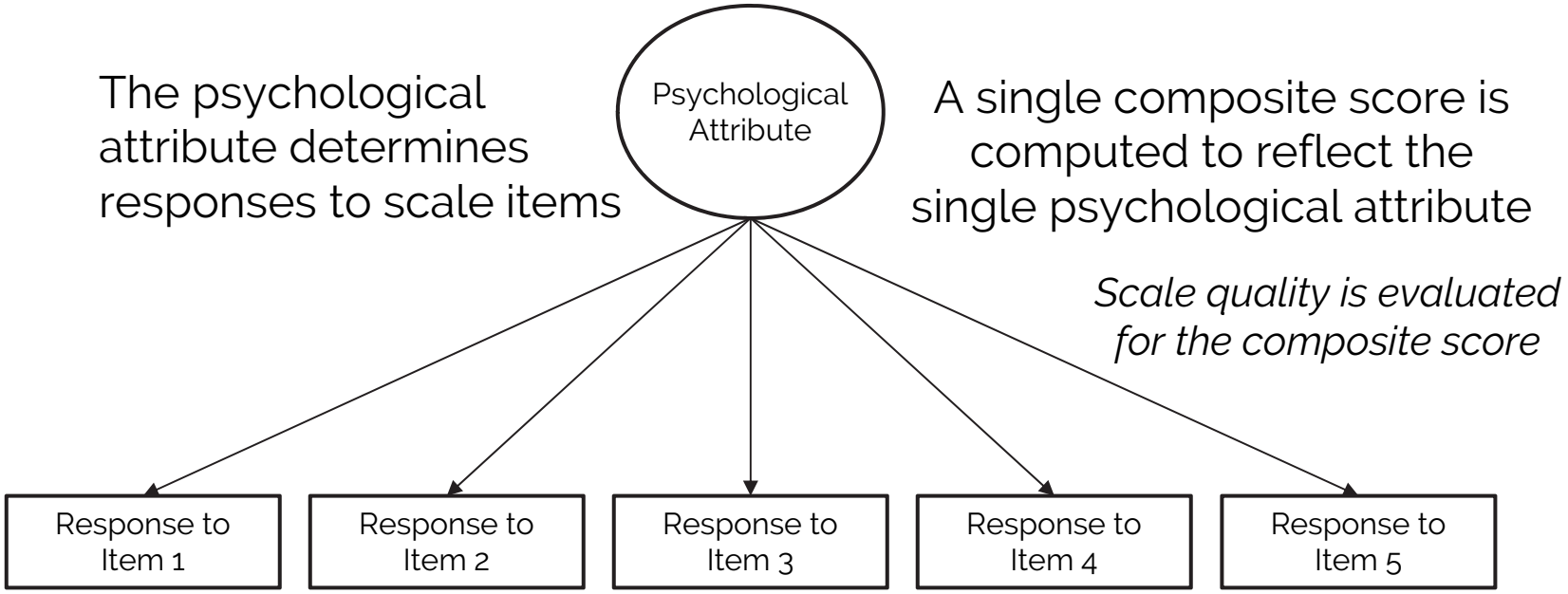


Furr & Bacharach (2014)

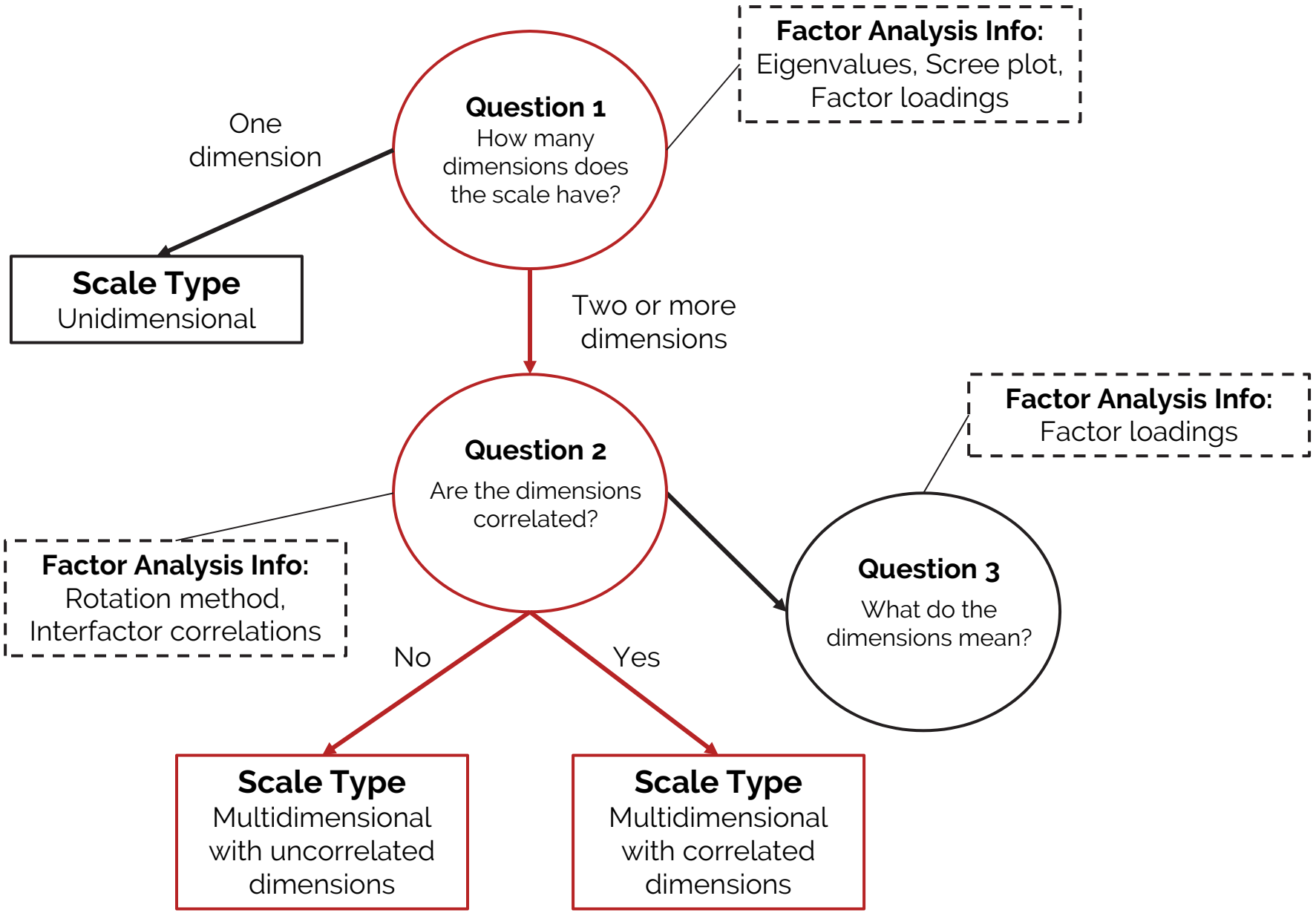
Unidimensional Scales

Unidimensional scales reflect a single psychological dimension

- For example, scores on a fatigue scale are interpreted as a measure of the amount of fatigue an operator is experiencing. This only makes sense if answers to scale items truly reflect fatigue and *only* fatigue.



Types of Scales



Furr & Bacharach (2014)

Multidimensional Scales

Multidimensional scales reflect two or more psychological attributes

- For example, a scale that measures personality may be comprised of 2 dimensions – an extraversion dimension and an openness to experience dimension

Dimensions can be correlated or uncorrelated

- Scales with correlated dimensions are called *scales with higher-order factors*
- Scales with uncorrelated dimension are called *scales with uncorrelated dimensions* (surprise!)

Scales with Higher-Order Factors

These scales include clusters of items that assess different psychological attributes called subscales

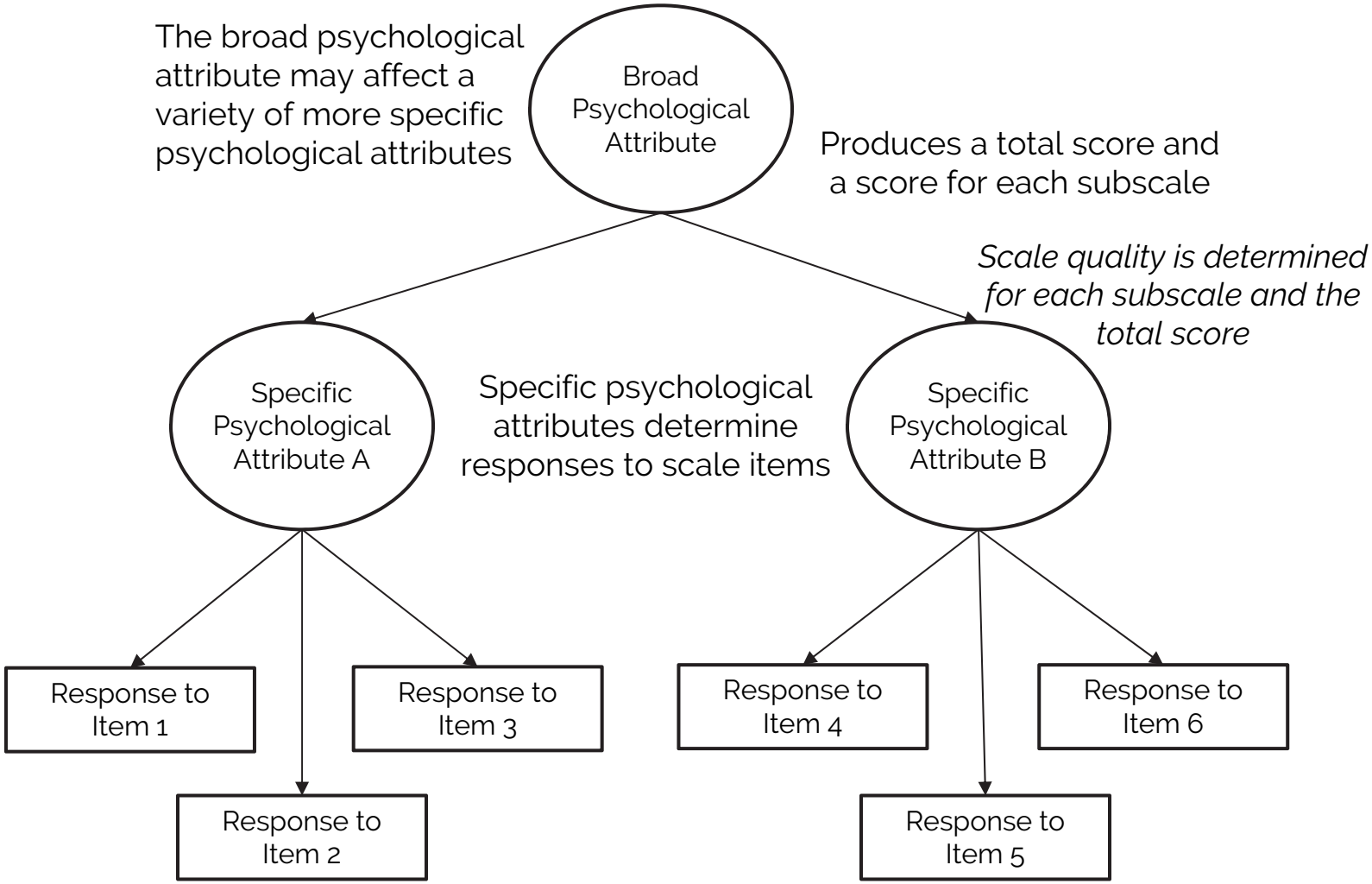
Each subscale reflects a different aspect of a broader psychological attribute and is itself unidimensional

- For example, the NASA-TLX includes 6 subscales that measure different aspects of workload. These subscales include mental, physical, and temporal workload as well as frustration, effort, and perceived performance

Because the subscales (dimensions) are correlated, these scales can produce various scores

- A score for each subscale
- A total score, combined across subscales

Scales with Higher-Order Factors



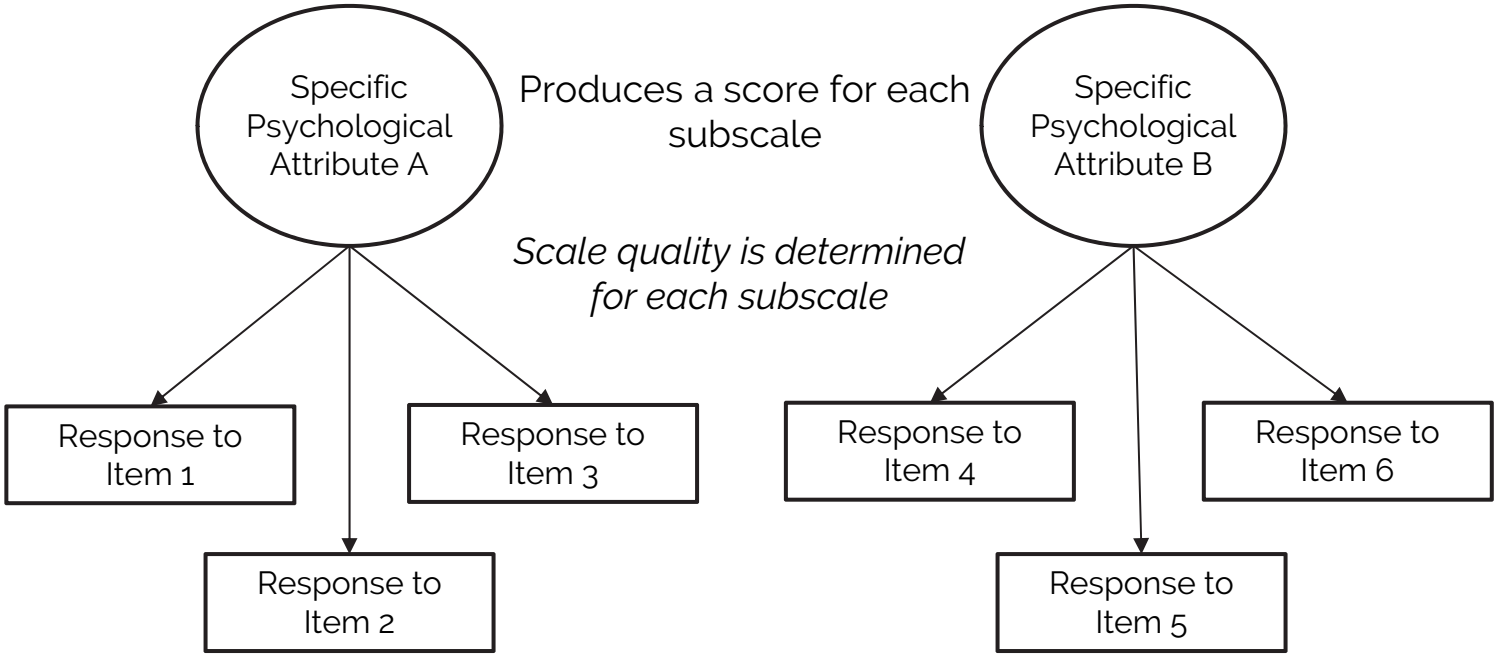
Scales with Uncorrelated Dimensions

These scales are similar to scales with higher-order factors except the subscales are not linked by a broader psychological attribute.

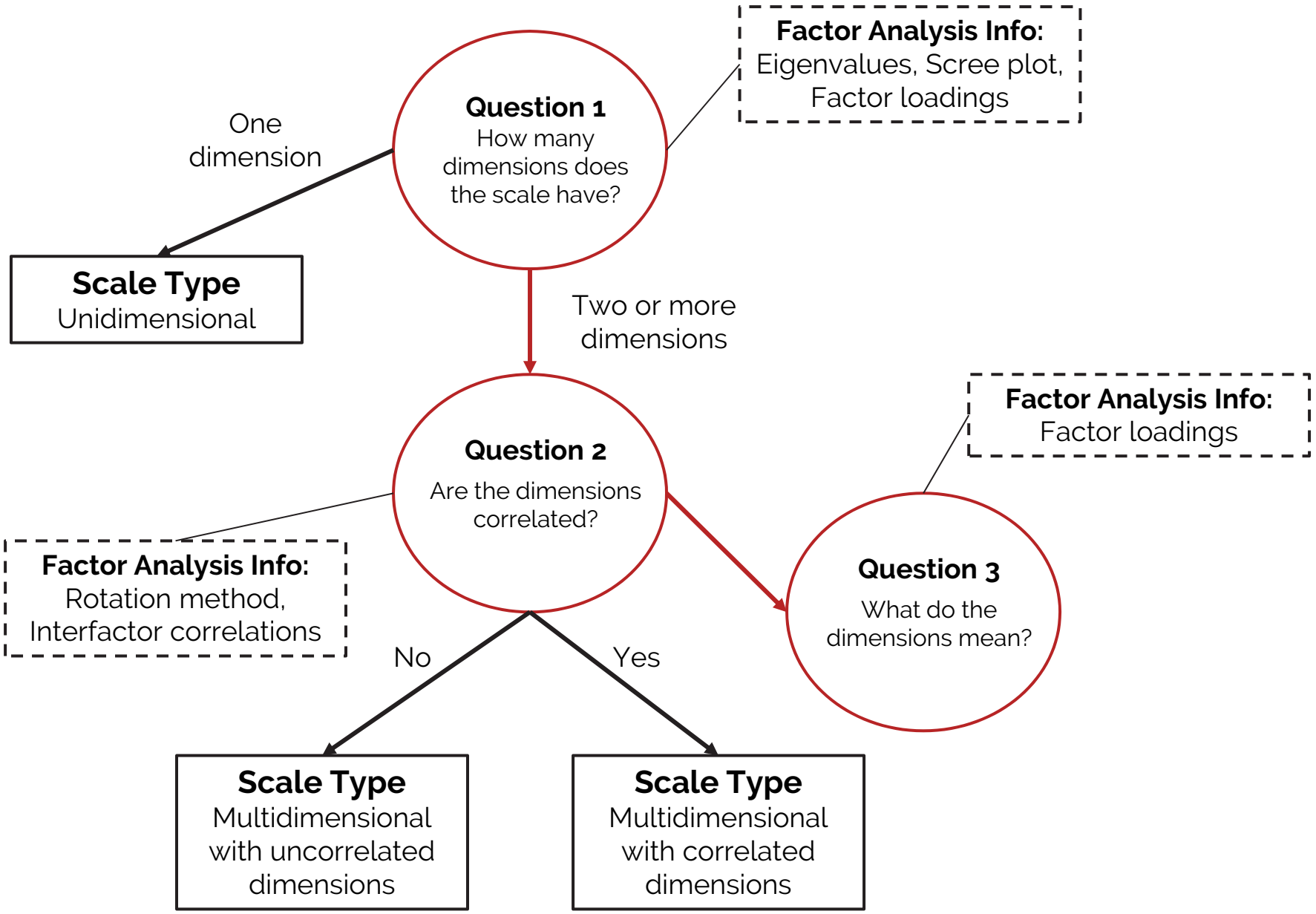
- Composite scores are computed for each subscale, but a total score should not be computed.

In essence, these scales are a set of unrelated unidimensional scales that are presented with their items mixed together

Scales with Uncorrelated Dimensions



Types of Scales



Furr & Bacharach (2014)

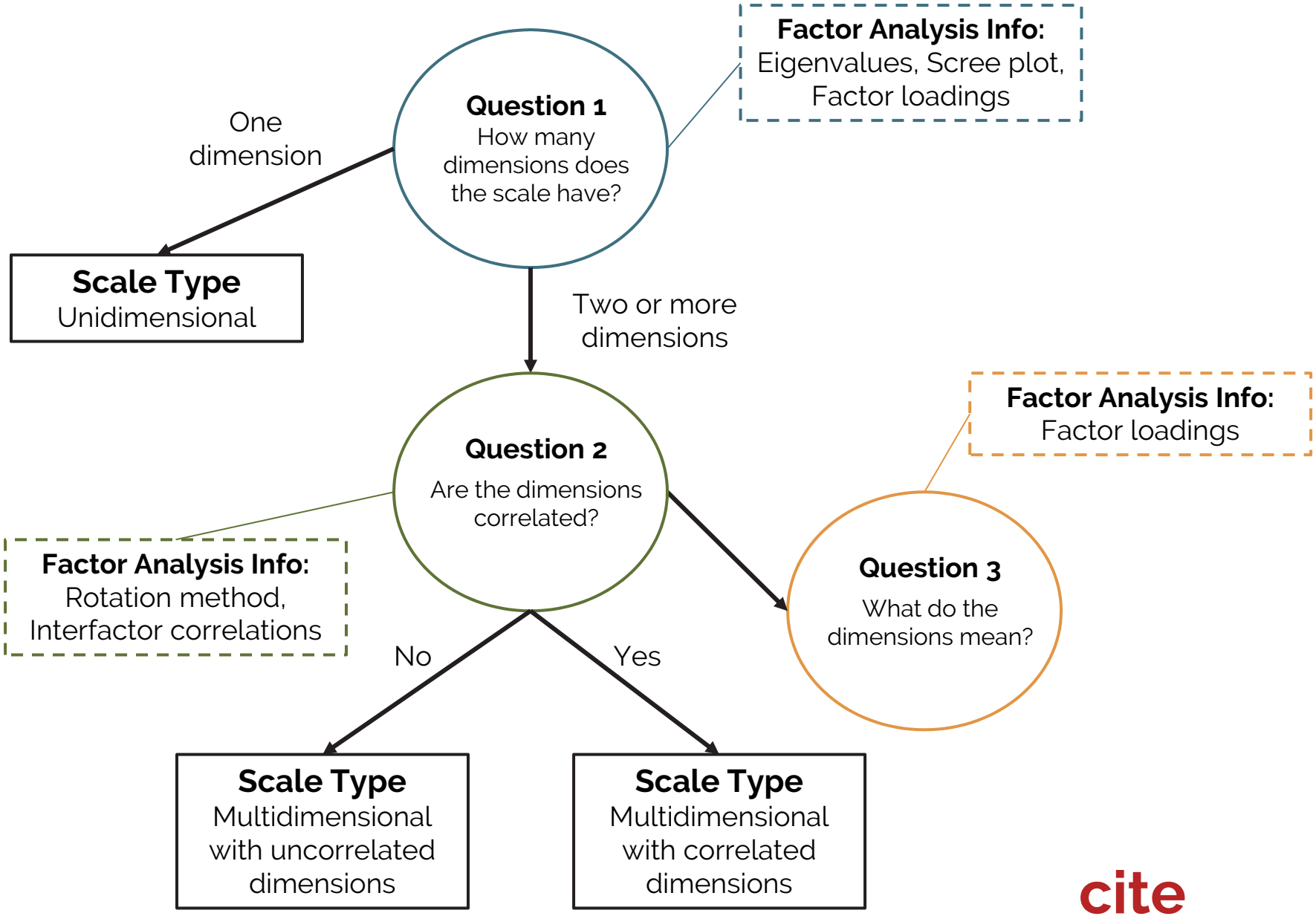
Psychological Meaning of Scale Dimensions

Researchers must conduct research that reveals the psychological attribute that is represented by each dimension

Factor analysis is a fundamental tool that researchers use in conducting this research

- Provides valuable information that helps answer the core questions of scale dimensionality

Types of Scales



Conceptual Overview of Factor Analysis

Factor analysis is the most common method for evaluating scale dimensionality

- Other statistical methods (e.g., cluster analysis, multidimensional scaling) are also available

Two types of factor analysis

- **Exploratory Factor Analysis (EFA)**
- Confirmatory Factor Analysis (CFA)

This method grounds clusters of items in empirical data rather than idiosyncratic interpretations

- Recall the personality scale that could be interpreted as having 1, 2, or 3 subscales

Factor Analysis Example

Imagine 100 soldiers rated how well 6 traits described them on the scale provided:

1. Talkative
2. Assertive
3. Imaginative
4. Creative
5. Outgoing
6. Intellectual

1	2	3	4	5
Completely unlike me	Somewhat unlike me	Neither like me nor unlike me	Somewhat like me	Completely like me

We can compute the correlations among the six items to help us identify and interpret the dimensions reflected

	Talkative	Assertive	Outgoing	Creative	Imaginative	Intellectual
Talkative	1.00					
Assertive	.66	1.00				
Outgoing	.54	.59	1.00			
Creative	.00	.00	.00	1.00		
Imaginative	.00	.00	.00	.46	1.00	
Intellectual	.00	.00	.00	.57	.72	1.00

Factor Analysis Example

Examining correlations is a very basic factor analysis

- Not typically possible with real data because there are more items and the correlational structure is less obvious

EFA simplifies this process

- Often an iterative process
- Results of one step lead researchers to reevaluate prior steps

Conducting Factor Analysis (Basics)

Input participants raw scores into a statistical software package (JMP, R, SPSS)

- Ratings should be reverse scored if necessary before conducting EFA

Step 1: Choose an extraction method

- Specific statistical technique implemented
- **Options:** principal axis factoring (PAF), maximum likelihood (ML), and principal components analysis (PCA)

  Results are often similar. However, some experts recommend PAF over PCA (MacCallum & Strahan, 1999). ML is typically reserved for CFA

Conducting Factor Analysis (Basics)

Step 2: Identify the number of factors and extract them

Researchers typically rely on *eigenvalues*

Eigenvalues are a special set of scalars associated with a linear system of equations (i.e., a matrix equation) that are sometimes also known as characteristic roots, characteristic values (Hoffman and Kunze 1971), proper values, or latent roots (Marcus and Minc 1988, p. 144)

Don't worry! You need to understand how eigenvalues are used not *necessarily* what they are…

How to Use Eigenvalues

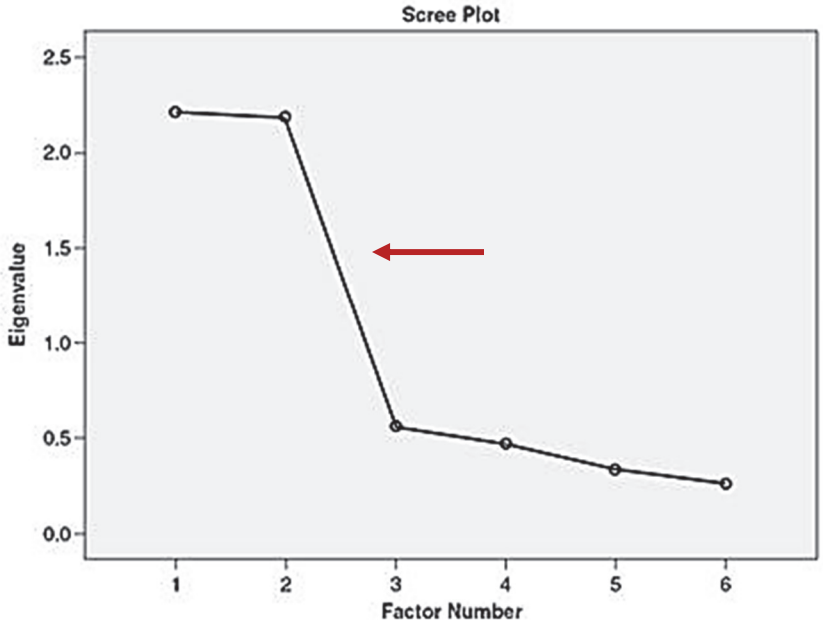
Examine the relative sizes of the eigenvalues

- Find point where all subsequent differences between values are relatively small
- The location of this point is indicative of the number of dimensions in the scale
- This same logic can be applied to scree plots

Total Variance Explained

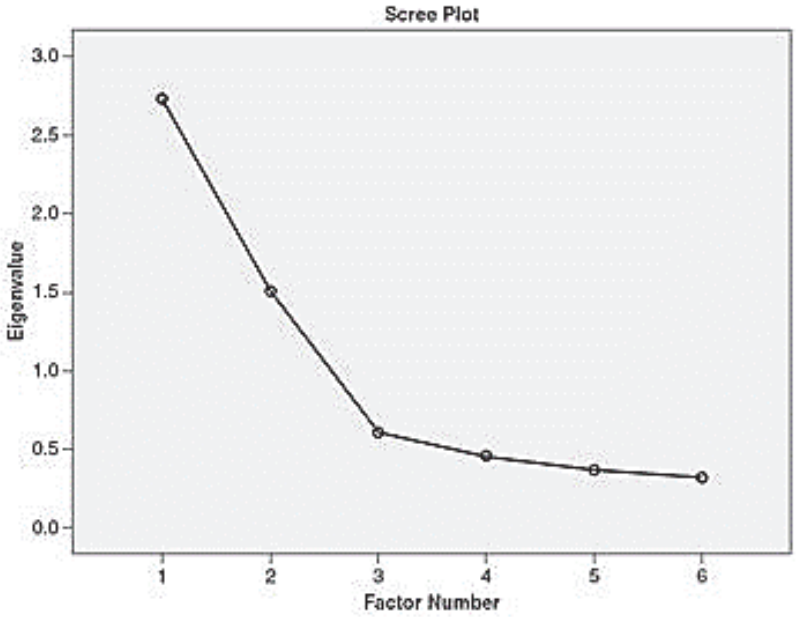
Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.195	36.578	36.578	1.836	30.599	30.599
2	2.173	36.222	72.800	1.808	30.131	60.730
3	.563	9.382	82.183			
4	.472	7.867	90.050			
5	.333	5.554	95.604			
6	.264	4.396	100.000			

Extraction Method: Principal Axis Factoring.



Caution

The appropriate number of factors is not always clear



1. Extract the number you think it is
2. Then examine the associations between items and that factor
3. Iterate if needed

If a clear number of dimensions cannot be determined, the scale likely needs to be revised

Conducting Factor Analysis (Basic)

Step 3: Decide how you will rotate factors

The purpose of this step is to clarify the psychological meaning of the factors

Two types of rotations

- **Orthogonal:** generates uncorrelated factors ("Varimax")
- **Oblique:** generates correlated or uncorrelated factors ("Promax", "Direct Oblimin")

There is rarely a reason for requiring that factors be correlated or uncorrelated during scale development

- Researchers typically select oblique rotations

Conducting Factor Analysis (Basic)

Step 4: Examine Item-Factor Associations

These associations are determined using *factor loadings*

- Each item has a loading on each factor

Examine the loadings to identify which items are most strongly linked to each factor

- The similarities among items linked most strongly to a factor points to the factor's psychological meaning

Factor loadings range from -1 to 1

- Interpreted as correlations or standardized regression coefficients depending upon the rotation

Factor Loadings

Orthogonal rotations yield factor loadings that can be interpreted as correlations between each item and each factor

Oblique rotations yield 2 types of factor loadings

- **Pattern coefficients:** item-factor association, controlling for the correlation between factors
- **Structure coefficients:** simple item-factor correlations

Consider the size and direction of the loading

- Loadings above .30 are “reasonable”, above .70 are “strong”
- Interpret the direction like a correlation – a negative loading indicates that high scores on the item are associated with low scores on the underlying factor

Factor Loadings Example

Imagine we chose an oblique rotation for the personality scale and obtained the following output

Factor Matrix[a]

	Factor	
	1	2
Intellectual	.942	.000
Imaginative	.764	.000
Creative	.604	.000
Assertive	.000	.849
Talkative	.000	.777
Outgoing	.000	.695

Loadings before rotation is applied

Pattern Matrix[b]

	Factor	
	1	2
Intellectual	.942	.000
Imaginative	.764	.000
Creative	.604	.000
Assertive	.000	.849
Talkative	.000	.777
Outgoing	.000	.695

Pattern Coefficients

Structure Matrix

	Factor	
	1	2
Intellectual	.942	.000
Imaginative	.764	.000
Creative	.604	.000
Assertive	.000	.849
Talkative	.000	.777
Outgoing	.000	.695

Structure Coefficients

Results obtained from real data are rarely this tidy

Conducting Factor Analysis (Basic)

Step 5: Examine the association among factors

Oblique rotations allow factors to be correlated or uncorrelated

The degree of correlation among factors determines how to score the scale

Factor Correlation Matrix

Factor	1	2
1	1.000	.000
2	.000	1.000

How should this scale be scored?

We would create 2 subscales

One composite for each subscale

No total score!

Course Objectives

1. Identify psychological measurement's goals and challenges
2. Understand basic measurement concepts and how they apply to psychological measurement
3. Understand scale development basics
4. Understand the importance of reliability and validity testing scales, factors that affect reliability and validity, and how to conduct reliability and validity testing

This page intentionally left blank.

Scale Quality

In applied contexts, researchers strive to make decisions about people, scale scores inform those decisions

The instruments that we use to make such decisions need to reliable and valid

> **Reliability:** extent to which scale scores are a function of respondents' true psychological differences as opposed to measurement error
>
> **Validity:** extent to which scale scores reflect what the scale is intended to measure

There are statistical methods for evaluating reliability and validity

Methods for Establishing Reliability

There is no single method that provides completely accurate estimates of reliability under all conditions

There are three primary methods for estimating reliability

- Alternate forms reliability
- Test-retest reliability
- Internal consistency reliability

Note: all of these methods are derived from the notion of parallel tests

Provide estimates of the proportion of observed score variance that is attributable to true score variance

$$\boldsymbol{Observed\ score\ = True\ score + Error}$$

Differ by the kind of data available and underlying assumptions

Alternate Forms Reliability

1. Develop two *parallel* versions of a scale
   - Items must probe the same psychological attribute
   - Equivalent amount of error variance

$$\textbf{Observed score} = \textbf{True score} + \textbf{Error}$$

2. Administer the two versions of the scale to the same group of people

3. Measure the correlation between scores on both versions

Potential Issues

We can never be **certain** that versions are parallel

- Versions considered "close enough" if they have similar means and standard deviations

Repeated testing may inflate correlations

Test-Retest Reliability

Avoids the parallel versions problem

1. Administer the same scale on two different occasions
   - Assume true scores are stable across the two occasions

$$\textbf{Observed score} = \textbf{True score} + \textbf{Error}$$

2. Measure the correlation between scale scores

Potential Issues

Some psychological attributes are more stable than others
- For example, mood vs. temperament

Take care to ensure that test conditions are as similar as possible and that test occasions are not too far apart
- Typically occur within 1-2 weeks

Internal Consistency Reliability

Doesn't require 2 versions of a scale or 2 test occasions

Can only be used with multi-item scales

Method differs slightly depending upon data type (binary vs. continuous) and desired estimation procedure (use of item variances, inter-item correlations, inter-item correlations)

All procedures are a two-step process

1. Administer scale to a group of people
   - Each item is treated as different "versions" of a scale
2. Estimate consistency of items using an equation

Most common procedure is Cronbach's alpha

Cronbach's alpha

1. Administer the scale to a group of people

2. Calculate the covariance between each pair of items
   - Covariance reflects the degree of association between 2 variables (items)
   - We hope to find that items in a scale positively covary

3. Sum the covariances
   - The larger the sum is, the more consistent the items are with each other

4. Submit the variance of scores on the complete test and the sum of the covariances into the following equation:

$$\propto = \textbf{estimated } R_{xx} = \left(\frac{k}{k-1}\right)\left(\frac{\sum c_{ii}}{s_x^2}\right)$$

sum of covariances

items

variance of total score

Cronbach's alpha

Produces a score between 0 and 1

The closer the score is to 1, the greater the internal consistency

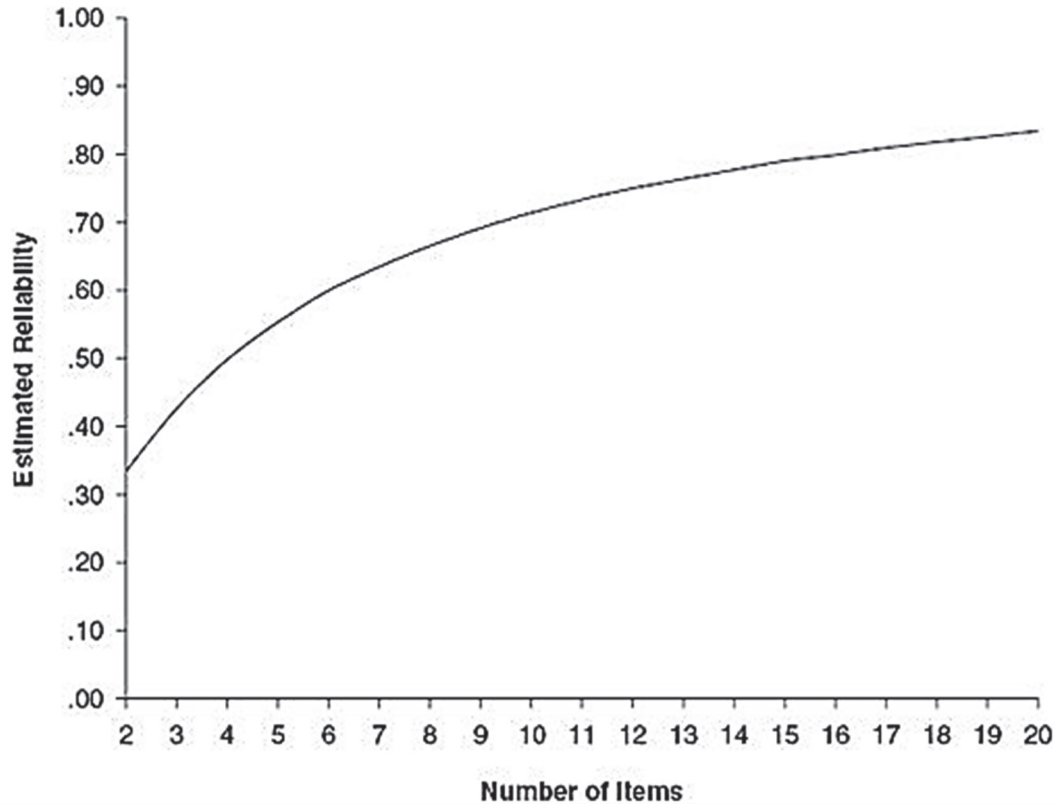
- $\alpha = 0.70$ is typically recognized as an "acceptable" level of internal consistency

Most statistical software also computes "Cronbach's alpha if deleted"

- Useful for identifying items that degrade internal consistency

Evidence suggests that Cronbach's alpha serves as a sort of "lower bound" on internal consistency

Note: Longer scales are more reliable



For a scale with an average inter-item correlation of 0.30

Conceptualizing Validity

Scale items are neither valid nor invalid. Researchers interpretations of those scores are valid or invalid

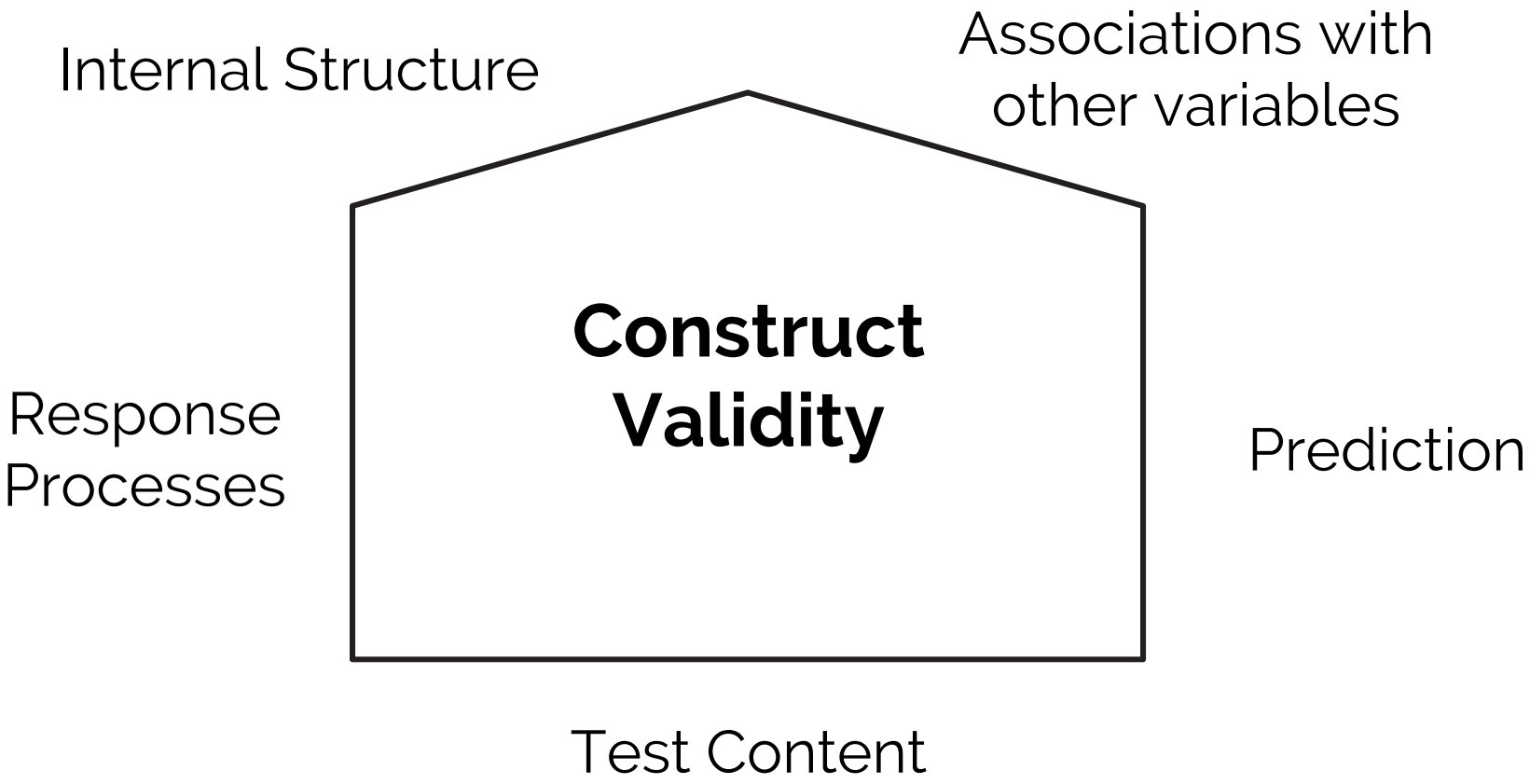
- Scales are like hammers. Someone might tell you a hammer is useful, but its usefulness depends on the task you are trying to complete

Validity is a matter of degree

Validity is a based on *theory* and *evidence* **NOT** someone's "experience"

Psychological theory AND empirical evidence are required to establish that a particular interpretation of scale scores is valid

Information Needed to Establish Validity



Methods for Establishing Validity

The **internal structure** (number of dimensions) of the scale should match the theoretically based structure of the scale

- Factor analysis! Speaks to the number of dimensions and how individual items map to these dimensions

The **psychological processes** that respondents actually use when responding to scale items should match the processes they should use

- Qualitative procedure called *Cognitive Interviewing*

Demonstrate **associations** between the scale and measures of theoretically related psychological attributes

- Commonly called *Convergent Validity*
- Statistical Procedures: Correlation, Regression

Methods for Establishing Validity

Demonstrate that the scale **predicts** behaviors that it should theoretically be able to predict

- Commonly called *Predictive Validity*
- Statistical Procedures: Regression

Course Objectives

1. Identify psychological measurement's goals and challenges
2. Understand basic measurement concepts and how they apply to psychological measurement
3. Understand scale development basics
4. Understand the importance of reliability and validity testing scales, factors that affect reliability and validity, and how to conduct reliability and validity testing

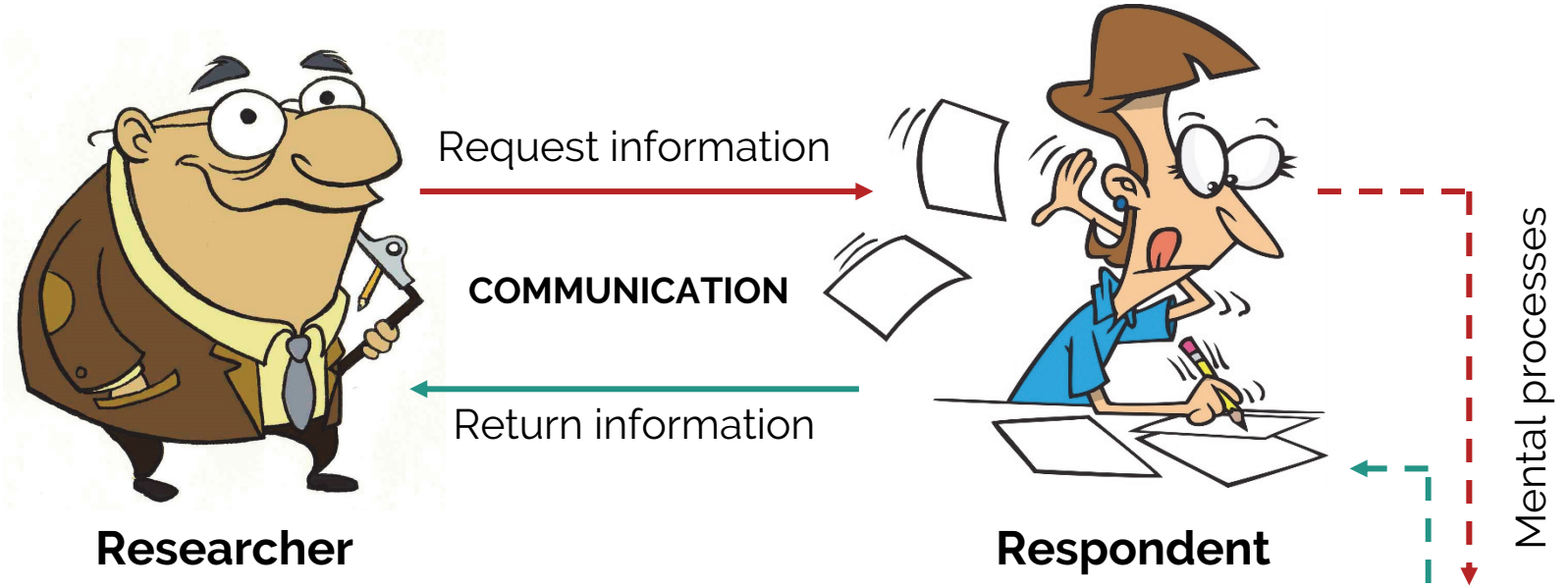
Questions?

This page intentionally left blank.

Back Up Slides

This page intentionally left blank.

Surveys are Conversations



Measurement error can occur at any of these steps

Process for Answering Questions

1. Understand question
2. Recall information
3. Form judgment
4. Respond using the format provided.

We want to develop questions that facilitate this process!

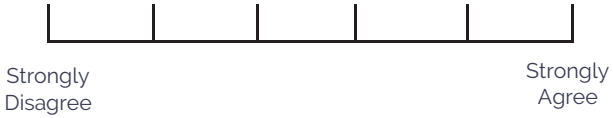
Question Types

Structured Questions

The researcher constrains how respondents answer the question

For example:

Scale Response



Quantitative analysis

Most of your survey questions

Useful for **measuring** psychological attributes

Unstructured Questions

The researcher doesn't constrain how respondents answer the question

For example:

Essay Response

Qualitative analysis

Most appropriate for interviews, but its okay to use sparingly in surveys

Useful for **understanding** unanticipated events and **discovering** problems

Structured Questions

Structured questions can take many forms

Dichotomous

☒ No

☐ Yes

Response-option format that forces respondents to select on of two options.

Multiple Choice

☐ Active Duty

☐ National Guard

☐ Civilian

☐ Contractor

Response-option format that forces respondents to select among several options.

Rank

1 Killer Robots

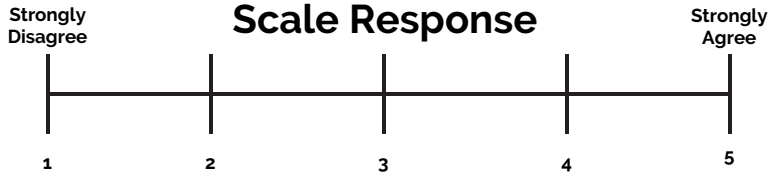
2 Aliens

4 Zombies

3 Vampires

Response-option format that forces respondents to rank order their preferences.

Scale Response



Response-option format that forces respondents to place their preferences on a scale.

How to Begin

Define the survey structure before beginning to generate specific questions

1. Write down each research question
2. Order the research questions logically
3. Define the type of information needed to answer each research question

For example:

Research question: Is the user interface for system X suitable?

Information: Measure task performance and collect ratings from operators across test conditions for the following constructs:

separate scales →
1. Usability of interface
2. Workload during task completion
3. Trust in system feedback

Scale Development

Begin developing questions to fill out each scale

- Psychological attributes are multi-dimensional
- Develop enough questions to address each dimension

> For example, there are various dimensions of workload – mental, physical, temporal (etc.). Measuring "workload" means that you must develop questions to reflect these different dimensions.

Be sure to include instructions to the larger survey and to each separate scale as needed

Item Writing Tips

Only ask one question at a time

Bad question: The interface was visually pleasing and easy to use.

Good question: The interface was easy to use.

Questions should be clear, concise, and grammatically simple

Bad question: If the satellite stops working, would you consider or not consider using X system to improve your ability to communicate with others?

Good question: If the satellite stops working, would you use system X to communicate?

Questions shouldn't lead respondents to a specific answer

Bad question: Shouldn't experienced operators be able to identify the target easily?

Good question: How easy was it to identify the target?

Item Writing Tips

Avoid using “loaded” or emotive language

Bad question: Do you believe the vehicle will protect soldiers from being maimed?

Good question: How helpful was the vehicle in accomplishing task X?

Questions should avoid absolutes and extremes

Bad question: Do you always create unique passwords for each system?

Good question: Do you create unique passwords for each system?

Avoid asking respondents to do unnecessary computations

Bad question: How many hours have you spent operating the system?

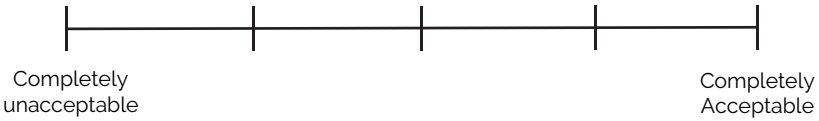
Good question: How many months have you spent operating the system?

Writing Response-Options

The response option should clearly match the item

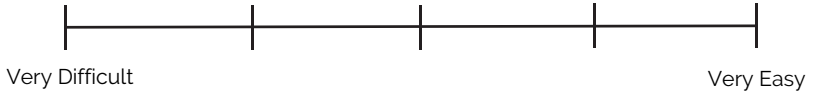
Bad Question

How easy was the interface to use?



Good Question

How easy was the interface to use?



Use 5- to 7-point scales when possible

Use balanced scales

Bad Question

How easy was the interface to use?

Good Question

How easy was the interface to use?

Keep format of questions as consistent as possible