INSTITUTE FOR DEFENSE ANALYSES

Vetting Custom Scales - Understanding Reliability, Validity, and Dimensionality

Dr. Heather Wojton, Project Leader

Dr. Stephanie Lane

December 2018

Approved for public release.
Distribution is unlimited.

IDA Non-Standard Document
NS D-9168

Log: H 2018-000279

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

IDA *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
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Task 2299(90), "Test Science Applications," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
Document review was done by Mr. Robert R. Soule and Dr. Heather Wojton from the Operational Evaluation Division.

For more information:
Dr. Heather Wojton, Project Leader
hwojton@ida.org • (703) 845-6811

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Non-Standard Document NS-9168

Vetting Custom Scales - Understanding Reliability, Validity, and Dimensionality

Dr. Heather Wojton, Project Leader

Dr. Stephanie Lane

Executive Summary

Within operational testing, we frequently want to measure some aspect of human system interaction (HSI) as part of operational effectiveness or operational suitability. We are often able to measure these HSI concepts with existing surveys from academic literature. These surveys, such as the NASA-TLX and the System Usability Scale (SUS), have already gone through extensive psychometric testing to demontrate that they are both reliable and valid.

However, there are situations in which analysts may need to create a custom survey, or scale, in order to evaluate an HSI concept. Motivating examples include (1) creating a new scale to account for shortcomings in a historical scale, (2) creating a new scale to measure a relatively new concept for which an existing scale may not exist, and (3) creating a new scale because the operator population is substantially different than the intended population of an existing scale.

For these situations in which an empirically vetted scale does not exist or is not suitable, a custom scale may be created. This document presents a comprehensive process for establishing the defensible use of a custom scale.

This document describes the full lifecycle of the scale validation process. At the highest level, this process encompasses (1) establishing validity of the scale, (2) establishing reliability of the scale, and (3) assessing dimensionality, whether intended or unintended, of the scale.

First, the concept of validity is described, including how validity may be established using operators and subject matter experts. The concept of scale reliability is also described, with guidelines for computing, interpreting, and using results to inform potential modifications to a custom scale. Next, exploratory factor analysis, or a method for investigating the dimensionality of a scale, is described along with a walkthrough of software implementation and results. Finally, confirmatory factor analysis, a technique for testing a priori hypotheses about dimensionality, is presented.

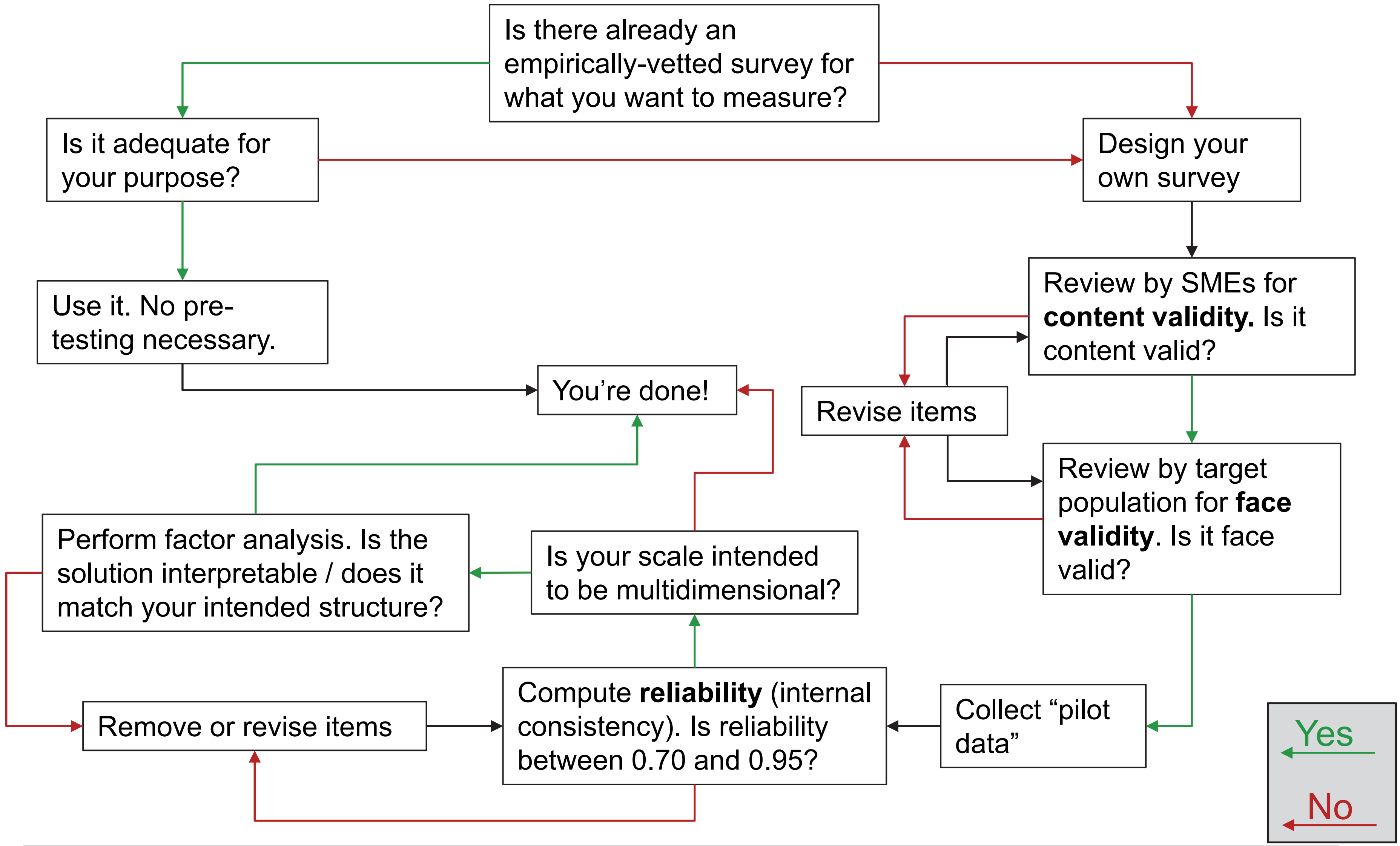
Vetting custom scales: Understanding reliability, validity, and dimensionality

Dr. Stephanie Lane

Roadmap

- The goal of measurement
- Establishing the validity of a survey
- Establishing the reliability of a survey
- Evaluating scale dimensionality

Roadmap to custom survey success



Establishing vocabulary

- **Item** – an individual question on a scale
- **Response option** – the number associated with the response
- **Unidimensional** – a scale reflects one (and only one) underlying concept
- **Multidimensional** – a scale reflects more than one concept. These concepts may or may not be related to each other.
- **Correlation** – the extent to which two variables are linearly related to each other
  - Important for establishing scale reliability

Correlation

Used to determine whether a **relationship** exists between two variables that are measured on an interval or ratio scale.

- Answers the questions:
  - Are two variables **linearly** related?
  - How strong is the relationship?

- In the panel figures to the right:
  - (a) might represent the correlation between height and weight
  - (b) might represent the correlation between the weight of a car and its MPG
  - (c) might represent the correlation between moon phase and crime rate
  - (d) might represent the correlation between anxiety and performance on a test

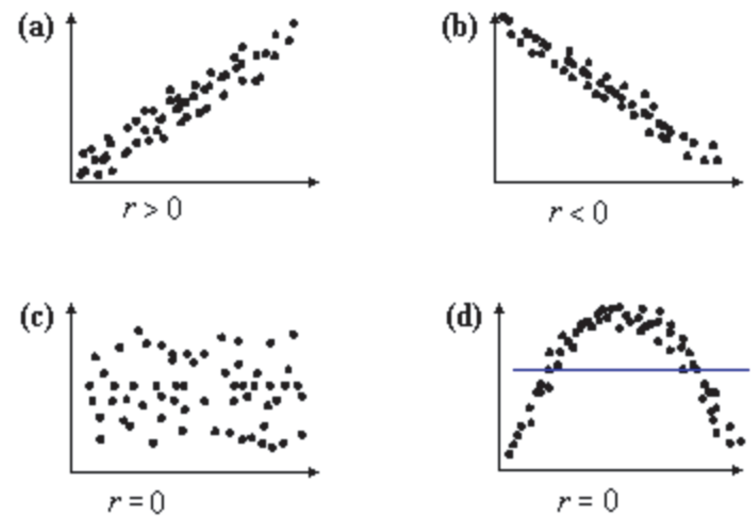


Figure taken from https://onlinecourses.science.psu.edu/stat500/node/60

Keep in mind that correlation is not causation

There is a positive correlation between ice cream sales and shark attacks.

Do ice cream sales cause shark attacks?

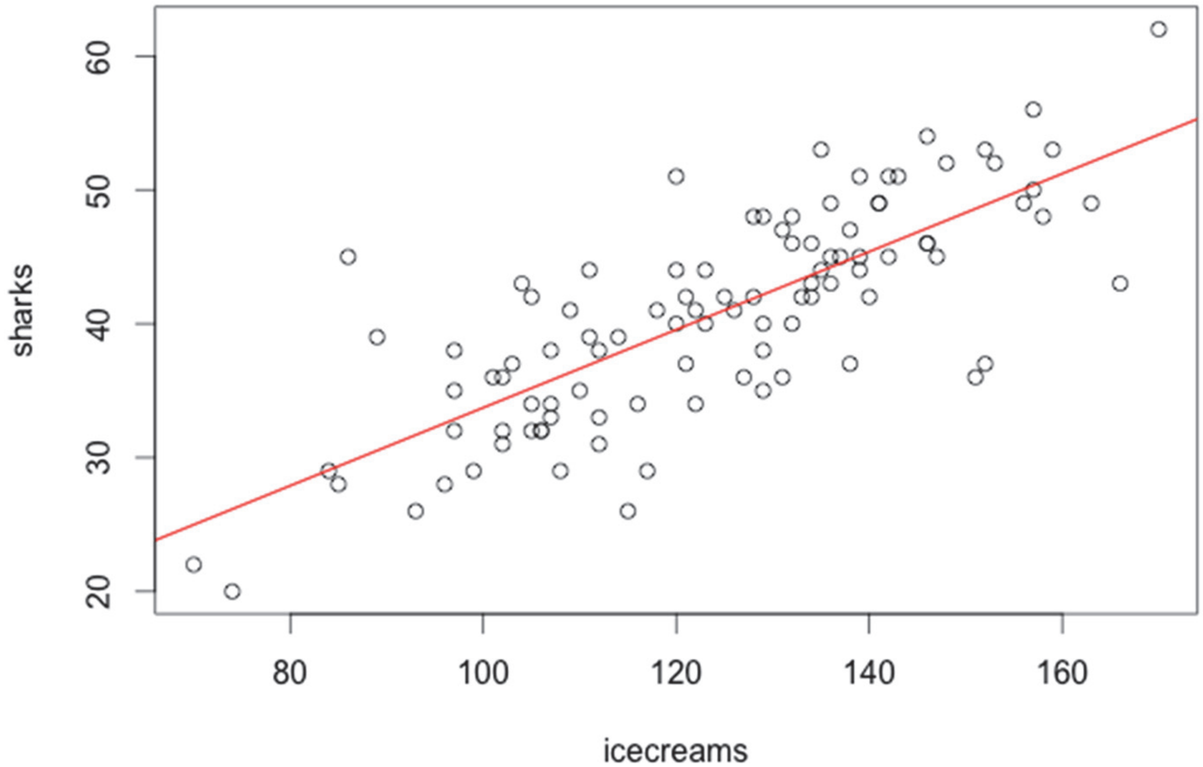


Figure taken from http://aittalam.github.io/assets/sharks/ice-sharks-fit.png

Thinking about measurement

- We frequently develop "measures" when we want to quantify some ability, concept, or experience
- Many examples of this exist in everyday life
  - Use inches to measure height
  - Use reaction time to measure cognitive ability
  - Use college first-year retention rate to measure student satisfaction
- Definitions of measurement:
  - Wikipedia*: "Measurement is the assignment of a number to a characteristic of an object or event."*
  - Stanley Smith Stevens: *"Measurement is the assignment of numerals to objects or events according to some rule."*
  - Research Methods in Psychology [Textbook]: *"Measurement is the assignment of scores to individuals so that scores represent some characteristic of the individuals."*

Common thread: systematic and quantitative

Applied measurement

We use measurement to quantify concepts that are difficult to measure directly, such as workload, intelligence, physical ability, etc.

For these measures to be used responsibly, we must establish that they are **valid** and **reliable** measures of our concept of interest.

A push-up test as a measure of physical performance

MILITARY MEDICINE, 181, 2:167, 2016

Inter-Rater Reliability and Intra-Rater Reliability of Assessing the 2-Minute Push-Up Test

Lynn Fielitz, PhD; Jeffrey Coelho, EdD; Thomas Horne, PhD; William Brechue, PhD

ABSTRACT The purpose of this study was to assess inter-rater reliability and intra-rater reliability of the 2-minute, 90° push-up test as utilized in the Army Physical Fitness Test. Analysis of rater assessment reliability included both total score agreement and agreement across individual push-up repetitions. This study utilized 8 Raters who assessed 15 different videotaped push-up performances over 4 iterations separated by a minimum of 1 week. The 15 push-up participants were videotaped during the semiannual Army Physical Fitness Test. Each Rater randomly viewed the 15 push-up and verbally responded with a "yes" or "no" to each push-up repetition. The data generated were analyzed using the Pearson product-moment correlation as well as the kappa, modified kappa and the intra-class correlation coefficient (3,1). An attribute agreement analysis was conducted to determine the percent of inter-rater and intra-rater agreement across individual push-ups.The results indicated that Raters varied a great deal in assessing push-ups. Over the 4 trials of 15 participants, the overall scores of the Raters varied between 3.0 and 35.7 push-ups. Post hoc comparisons found that there was significant increase in the grand mean of push-ups from trials 1–3 to trial 4 ($p < 0.05$). Also, there was a significant difference among raters over the 4 trials ($p < 0.05$). Pearson correlation coefficients for inter-rater and intra-rater reliability identified inter-rater reliability coefficients were between 0.10 and 0.97. Intra-rater coefficients were between 0.48 and 0.99. Intra-rater agreement for individual push-up repetitions ranged from 41.8% to 84.8%. The results indicated that the raters failed to assess the same push-up repetition with the same score (below 70% agreement) as well as failed to agree when viewed between raters (29%). Interestingly, as previously mentioned, scores on trial 4 increased significantly which might have been caused by rater drift or that the Raters did not maintain the push-up standard over the trials. It does appear that the final push-up scores received by each participant was a close approximation of actual performance (within 65%) but when assessing physical performance for retention in the Army, a more reliable test might be considered.

A drinking survey as a measure of health risk

0145-6008/03/2705-0826$03.00/0
ALCOHOLISM: CLINICAL AND EXPERIMENTAL RESEARCH

Vol. 27, No. 5
May 2003

The Reliability and Validity of the Self-Reported Drinking Measures in the Army's Health Risk Appraisal Survey

Nicole S. Bell, Jeffrey O. Williams, Laura Senier, Shelley R. Strowman, and Paul J. Amoroso

Background: The reliability and validity of self-reported drinking behaviors from the Army Health Risk Appraisal (HRA) survey are unknown.

Methods: We compared demographics and health experiences of those who completed the HRA with those who did not (1991–1998). We also evaluated the reliability and validity of eight HRA alcohol-related items, including the CAGE, weekly drinking quantity, and drinking and driving measures. We used Cohen's κ and Pearson's r to assess reliability and convergent validity. To assess criterion (predictive) validity, we used proportional hazards and logistical regression models predicting alcohol-related hospitalizations and alcohol-related separations from the Army, respectively.

Results: A total of 404,966 soldiers completed an HRA. No particular demographic group seems to be over- or underrepresented. Although few respondents skipped alcohol items, those who did tended to be older and of minority race. The alcohol items demonstrate a reasonable degree of reliability, with Cronbach's $\alpha = 0.69$ and test-retest reliability associations in the 0.75–0.80 range for most items over 2- to 30-day interims between surveys. The alcohol measures showed good criterion-related validity: those consuming more than 21 drinks per week were at 6 times the risk for subsequent alcohol-related hospitalization versus those who abstained from drinking (hazard ratio, 6.36; 95% confidence interval=5.79, 6.99). Those who said their friends worried about their drinking were almost 5 times more likely to be discharged due to alcoholism (risk ratio, 4.9; 95% confidence interval=4.00, 6.04) and 6 times more likely to experience an alcohol-related hospitalization (hazard ratio, 6.24; 95% confidence interval=5.74, 6.77).

Conclusions: The Army's HRA alcohol items seem to elicit reliable and valid responses. Because HRAs contain identifiers, alcohol use can be linked with subsequent health and occupational outcomes, making the HRA a useful epidemiological research tool. Associations between perceived peer opinions of drinking and subsequent problems deserve further exploration.

Key Words: Alcohol, Military, Reliability, Validity, Survey.

A custom survey of organizational commitment

MILITARY PSYCHOLOGY, 2003, *15*(3), 191–207

The Measurement and Consequences of Military Organizational Commitment in Soldiers and Spouses

Paul A. Gade and Ronald B. Tiggle
U.S. Army Research Institute

Walter R. Schumm
Kansas State University

Based on the work of Meyer and Allen (1997), we derived a set of abbreviated scales to measure affective and continuance organizational commitment and conducted an extensive examination of the factor structure and reliability of these scales. The relation of these 2 abbreviated scales of organizational commitment to critical organizational outcomes was examined and tested. Results showed that affective and continuance commitment combined to influence subsequent soldier performance on job knowledge tests in opposite ways, suggesting a causal link between commitment and performance. Relations between affective and continuance commitment combinations and soldier-reported retention intentions, morale, and readiness were also explored. Scales developed to measure spouse commitment to the Army showed a factor structure that was comparable to that of soldiers and consistent with the dimensions of affective and continuance commitment.

Empirically vetting surveys

Surveys go through a vetting process that allows us to answer key questions:

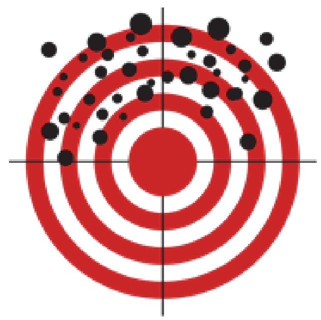
- Does my scale measure what I think it measures?

- Do the items that I want to measure the same thing *actually* measure the same thing?

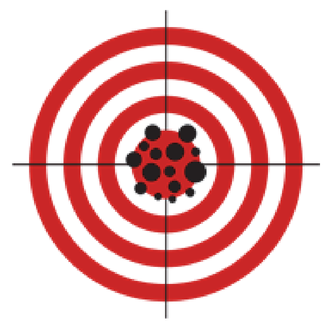
- Can I compute a composite score from my scale?

Establishing the **validity** and **reliability** of a scale enables us to answers these questions.

Key concepts within scale validation are reliability and validity



Not reliable, not valid



Reliable, not valid

Reliable, valid

A scale that is not reliable **cannot be valid.**

Thinking about reliability and validity in everyday measures

- Height in inches as a measure of weight
  – Reliable, not valid

- A clock set to Zulu time as a measure of the current time in Alexandria, Va.
  – Reliable, not valid

- Hunger level as a measure of stress
  – Not reliable, not valid

- Weight in pounds as a measure of weight
  – Reliable, valid

Establishing reliability is **necessary but not sufficient** for establishing validity

There are many types of validity – we will discuss four types.

- We will discuss four key types of validity to consider when creating your own custom scale.
  - Face validity
  - Content validity
  - Criterion validity
  - Construct validity

- Importantly, validity does not exist in a vacuum! The context and intended use of the scale are important when deciding whether or not a scale is valid.

Content validity is decided upon by experts

- Content validity relies on a subjective judgment, and is arguably the earliest stage of scale validation

- Subject matter experts decide whether a survey item is an appropriate measure of the **content** they intend to measure, and whether it has good **coverage** of the concept

- Key question answered by content validity – would independent subject matter experts, if shown your scale, agree that it fully reflects the concept of interest?

**Note: there is no reason to move forward with a scale that is not content valid. It should be revised before any collection of pilot data.*

Content validity: WWWWH

- Who?
  – Subject matter experts (SMEs), such as survey administrators
- What?
  – Do the survey items reflect what they are supposed to measure, and with sufficient breadth?
- When?
  – After the survey is written, but before pilot data is formally collected
- Why?
  – To establish that the survey sufficiently covers the relevant domain, and that the survey content reflects the intended concept
- How?
  – Panel of SMEs discuss adequacy of items

Content validity exercise

We create a short scale to assess **usability** of a new computer system. The entire system is new, including a non-QWERTY keyboard, a new ergonomic mouse, and a widescreen display.

Scale A:	Scale B:
The screen was easy to read.	The screen was easy to read.
The keyboard was easy to use.	The screen showed targets clearly.
The mouse was easy to use.	The screen showed text clearly.

Which scale demonstrates more **content validity** for assessing usability of the new system?

Scale A, because we want the usability of the entire system.

Establishing face validity of a scale is an important part of survey validation

- Establishing **face validity** is an early stage of scale validation.
- Face validity requires no computation. It requires relevant persons reviewing a scale to comment on whether the scale appears to measure what it is supposed to measure.
- Key question answered by face validity – would a member of the target population (e.g., operators) understand or recognize the question they are responding to?
- If a custom scale is not face valid, it should be revised.
  - There are instances when a scale that is not face valid could be useful. These instances will not be frequent in the context of operational testing.

Face validity: WWWWH

- Who?
  - Target population (e.g., operators)
- What?
  - Do the survey items *look like* what they are supposed to measure?
- When?
  - After the survey is written, but before pilot data is formally collected
- Why?
  - To establish that the audience reads the question as SMEs intend
- How?
  - Cognitive interviewing with relevant audience

Face validity exercise

Suppose we wanted to measure operator **trust** in a new system.

- **Face valid**: “I believe the output provided by the system.”

- **Not face valid**: “The buttons mask important visual cues.”

The second item is **not** face valid because an operator would not perceive that the item gauged their trust in the system.

Recapping face validity versus content validity

- Face validity is established by the **target audience** – the individuals who will be responding to your survey

- Content validity is established by **subject matter experts** – the individuals who are administering your survey

- **Both** types of validity should be established before pilot data are collected

Criterion validity – the relationship of the scale to other measures

- Criterion validity measures how your scale performs with reference to some other **criterion**
  - There are three types of criterion validity; they have different names depending on when you collect your scale versus your criterion data
    - » Predictive validity (your scale data are collected before your criterion)
    - » Concurrent validity (your scale data are collected at the same time as your criterion)
    - » Postdictive validity (your scale data are collected after your criterion)

- Key questions answered by criterion validity:
  - Does it relate to things it should relate to?
  - Does it NOT relate to things it shouldn't relate to?

- If your scale demonstrates unexpected relationships with other measures, careful thought should be given to the scale before moving forward with it in a future survey administration.

- How to assess: a simple correlation

Criterion validity – an example

The Scholastic Aptitude Test (SAT) is intended to be a measure of academic ability, and is administered to students nationwide for college admission.

Here, we see the substantial relationship between SAT critical reading and writing scores with first-year English grades in college.

We would say that the SAT has predictive validity, a form of criterion validity.

Figure 4. The relationship between SAT critical reading and writing scores and first-year English grades.[21]

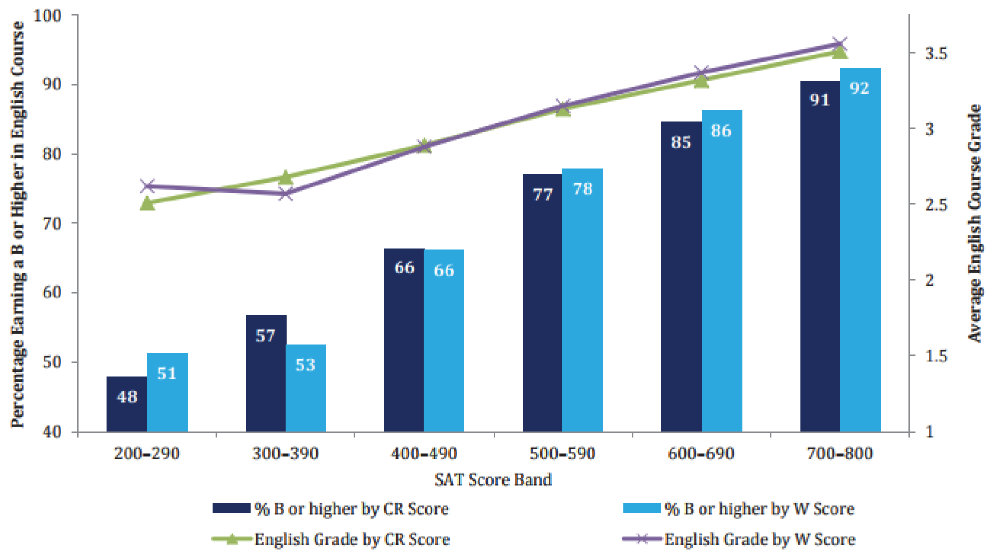


Figure 4 taken from https://research.collegeboard.org/sites/default/files/publications/2015/6/research-report-sat-validity-primer.pdf

Criterion validity: WWWWH

- Who?
  - Data analyst
- What?
  - Does my scale relate to things it should relate to?
- When?
  - After pilot data has been collected
- Why?
  - To help establish that the scale measures what you think it measures
- How?
  - Pearson correlation with other relevant metrics (e.g., other survey data, performance data, or behavioral data)

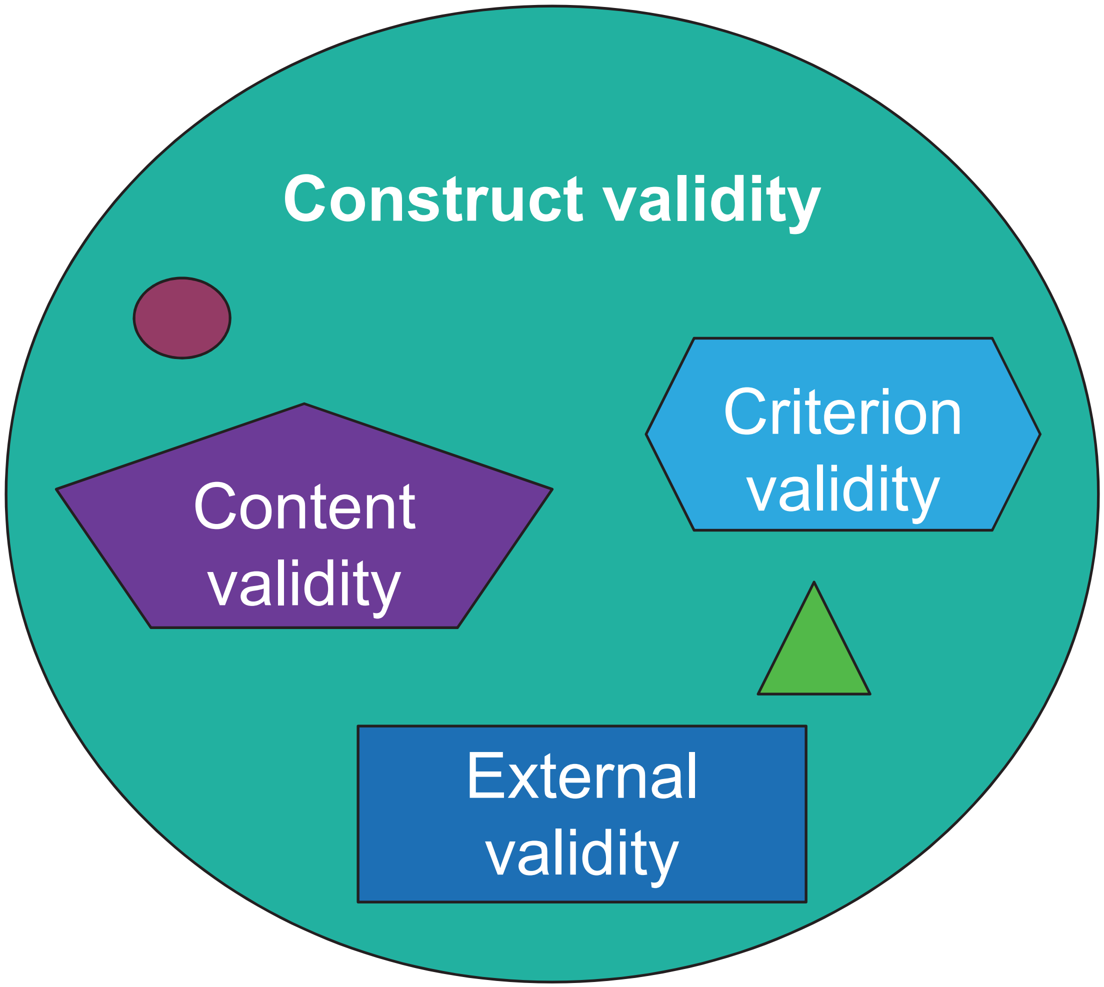
Criterion validity – how to execute in practice

- Suppose you create a new 10-item scale to measure **trust** in an automated system.

- To establish criterion validity, you could administer your custom **trust** scale. You could also ask one additional item asking about system **reliance**.

- You could then compute a correlation between system trust and system reliance. Trust and reliance should relate to each other. Establishing a correlation between trust and reliance would provide **concurrent validity** for your new trust scale.

Construct validity – our ultimate goal



Establishing validity of your scale is critical

- Establishing the validity of your custom scale is a critical step, and importantly, **only needs to happen once**

- If there is a concept (e.g., trust in an automated system) that needs to be measured over time, you can validate your scale once
  - Your scale can then be used in all future operational tests

- The validity of your scale only needs to be revisited if its **intended use** or **target audience** changes substantially

Establishing validity is consistent with DOT&E guidance to pre-test surveys

Pre-Testing Survey Instruments

Pretesting surveys (often referred to as pilot-testing) is a deliberate review of the survey to ensure that respondent answers will be what the testers need and are useful for the required analyses. Pretesting surveys should not be confused with the traditional pilot test – an event that immediately precedes the operational test to confirm that all data collection procedures are working properly. Pretesting should occur as part of the test planning process, prior to submitting the Operational Test Plan for DOT&E approval.

Pretesting is widely and strongly recommended by survey experts in academia, industry, and the military. The Army Research Institute's (ARI's) *Questionnaire Construction Manual* presents the most straightforward mandate for pretesting [emphasis in original]:

> <u>Pretesting is an important and essential procedure to follow before administering any questionnaire</u>.... Pretesting may seem to some uninformed individuals to be a waste of time, especially when the author may have asked several people in his/her office to critique the questions, or perhaps even asked a questionnaire specialist to critique it. However, <u>pretesting is an investment that is well worthwhile</u>. It is crucial if the decision that will result from the questionnaire is of any importance.

From Jan 6, 2017 memo, "Survey Pre-Testing and Administration in Operational Test and Evaluation."

Reliability

A related, important concept is scale reliability

There are many types of reliability. All types of reliability pertain to consistency.

Types of reliability

- One type measures the consistency of scores across multiple individuals (inter-rater reliability).

- Another type measures the consistency of a scale across multiple survey administrations (test-retest).

- Another type measures the consistency of items *within* a scale (internal consistency).

Inter-rater reliability

- **Inter-rater reliability** measures the agreement among a set of raters

- Key question answered by inter-rater reliability – is there agreement among my raters? If not, either the scale should be redesigned or the raters should be retrained.

- Several measures of inter-rater reliability exist. **Cohen's kappa** is a common measure of inter-rater reliability for two raters. **Fleiss' kappa** is a generalization that works for any number of raters. **Intra-class correlation** is another suitable measure of computing inter-rater reliability.

Cohen's kappa: definition

TABLE 1

An Agreement Matrix of Proportions

	Category	Judge A 1	2	3	p_{iB}
Judge B	1	.25 (.20)*	.13 (.15)	.12 (.15)	.50
	2	.12 (.12)	.02 (.09)	.16 (.09)	.30
	3	.03 (.08)	.15 (.06)	.02 (.06)	.20
	p_{iA}	.40	.30	.30	$\sum p_i = 1.00$

$$p_o = .25 + .02 + .02 = .29$$
$$p_c = .20 + .09 + .06 = .35$$

* Parenthetical values are proportions expected on the hypothesis of chance association, the joint probabilities of the marginal proportions.

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

- p_o is the relative agreement among raters and p_e is the probability of agreement due to chance
- It is interpreted as the "proportion of joint judgments in which there is agreement, after chance agreement is excluded."

Table and quotation from Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20 37-46.

Inter-rater reliability is an important consideration when humans are involved in assigning scores.

More commonly, we are interested in examining the consistency of scores across items.

A reliable scale will provide you with similar answers over time

- More similarity → more confidence that our scale is measuring a concept reliably → more confidence in our results

- Specifically, we have more confidence that a composite score we create (e.g., a roll-up calculation across multiple items) reflects the *true* score

- We can measure the consistency of items within a scale using a measure called **Cronbach's alpha**

Internal consistency is a measure of how well your items "hang together"

What is the extent to which my items measure a similar thing?

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{y_i}^2}{\sigma_x^2}\right)$$

- k is the number of items
- $\sigma_{y_i}^2$ is the variance of each item i
- σ_x^2 is the variance of the total score $(y_1 + y_2 + \ldots y_k)$

Yields a number between 0 and 1.

Helpful note:

In order to avoid confusion with the “alpha” we frequently refer to in operational testing…

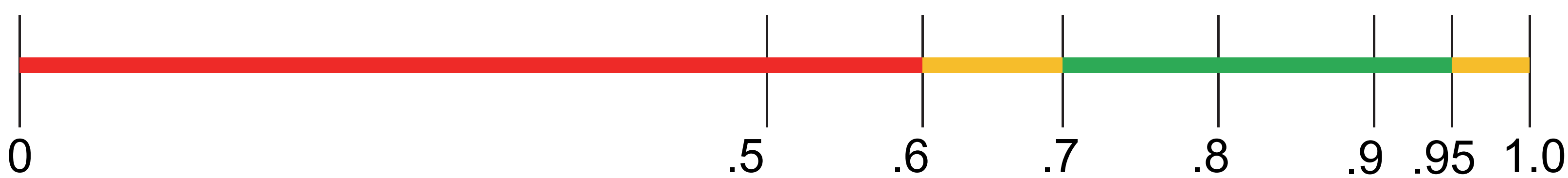
We always refer to this measure of internal consistency as “Cronbach’s alpha.” Never just “alpha.”

Cronbach's alpha: WWWWH

- Who?
  - Data analyst
- What?
  - Do the survey items interrelate highly with one another?
- When?
  - After the collection of pilot data
- Why?
  - To establish the items reliably measure the concept
- How?
  - Compute Cronbach's alpha in JMP or R (or other software)

We want internal consistency to be high… but not too high.

Cronbach's alpha	Internal consistency is judged to be
>= .9	Excellent
.8 - .9	Good
.7 - .8	Acceptable
.6 - .7	Questionable
.5 - .6	Poor
< .5	Unacceptable



A Cronbach's alpha > .95 indicates that your items are too consistent, and likely redundant.

We can use internal consistency to guide decisions

Suppose we have a scale measuring the performance of a lever.

We obtain the following results:

Cronbach's alpha for full scale = .78

Item	Cronbach's alpha if item dropped
The lever is strong.	.72
The lever is dependable.	.71
The lever is reliable.	.72
The lever behavior is predictable.	.73
The lever is easy to see from a distance.	.84

*Notional data used.

From Cronbach's alpha, we can obtain information necessary to:

- Drop poorly performing items
  - If an item harms your internal consistency, you may want to drop it or consider it separately
- Justify item groupings
  - If a set of items show good internal consistency (between .7 and .95), they are measuring a similar concept and could be aggregated
- Potentially create a short-form questionnaire
  - For example, a short form could be created for IOT&E based on data collected during an operational assessment or user test
    - » e.g., eight items could be reduced to the four best performing items for future survey administrations
  - Helpful in survey administration environments with minimal time to spare

How to improve scale reliability

- Use Cronbach's alpha as a tool
  - Inform which items to drop based on data

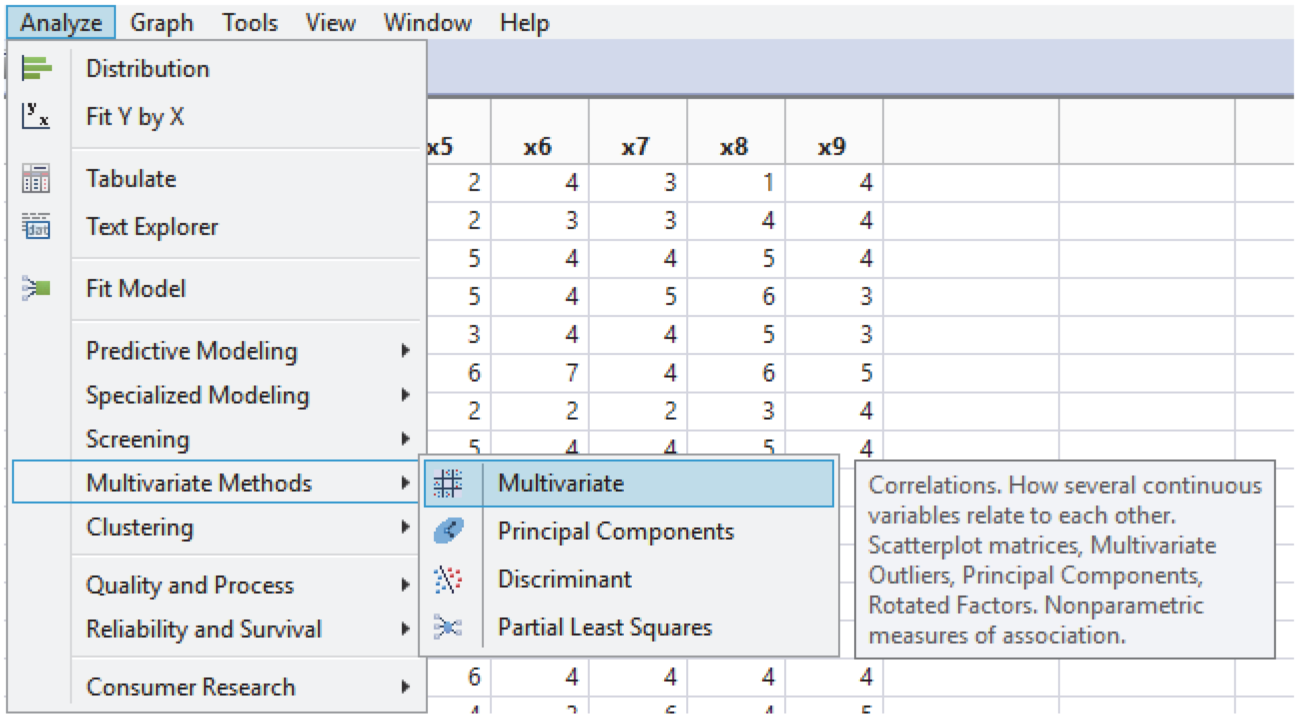
- Write better items
  - Screen any confusing items
  - Screen any double-barreled items

- Write more items
  - But manage respondent burden when lengthening scale

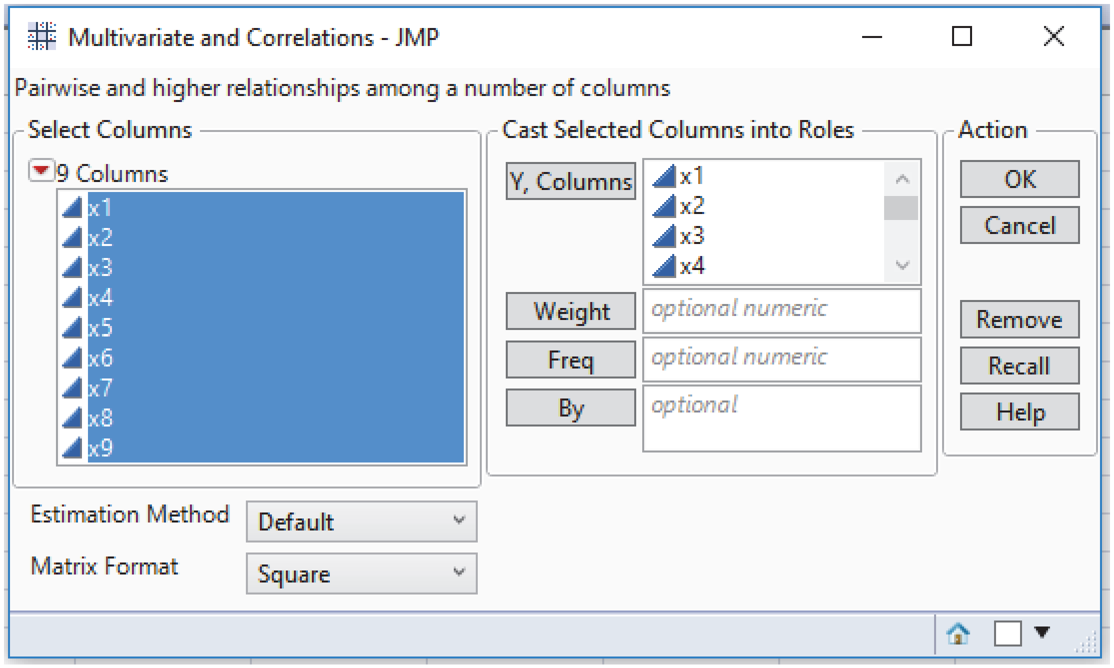
- Write items that will perform more consistently
  - For example, if we were measuring anger, the items "My anger sometimes interferes with my work," and "I punch a wall every time I am angry" would not perform consistently

Bottom line: Cronbach's alpha is an easy measure to compute for use in survey validation. It allows us to assess the similarity of items within a scale or subscale.

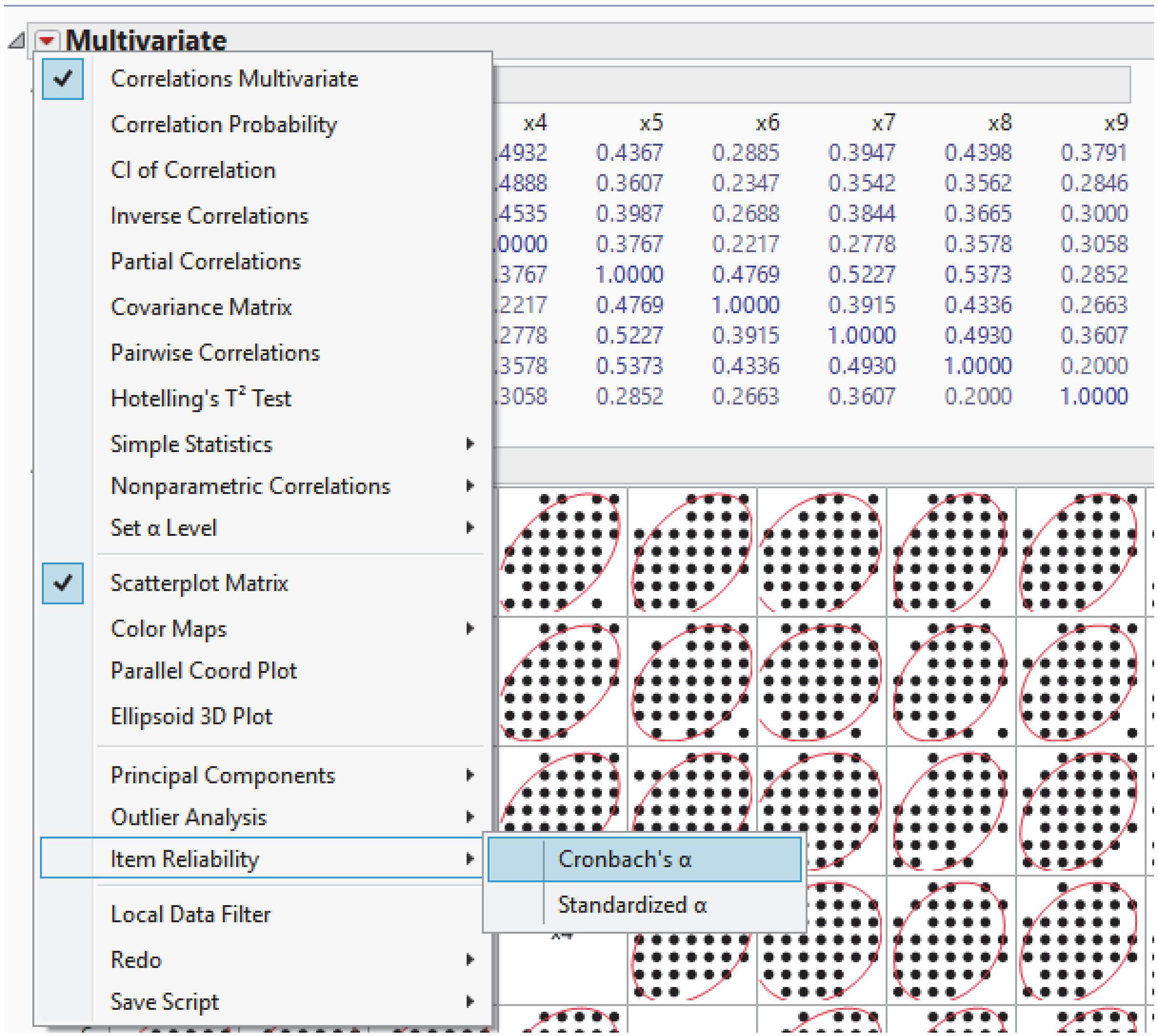
Computing Cronbach's alpha in JMP



Computing Cronbach's alpha in JMP



Computing Cronbach's alpha in JMP



Computing Cronbach's alpha in JMP

Cronbach's α

	α
Entire set	0.8479

Excluded Col	α
x1	0.8224
x2	0.8316
x3	0.8279
x4	0.8340
x5	0.8245
x6	0.8433
x7	0.8300
x8	0.8290
x9	0.8445

What is Cronbach's alpha NOT?

- It is **not** a formal test of dimensionality.
- What if you've written a scale that should have two or three distinct, but potentially related sub-scales?
- You will need to explicitly examine item dimensionality.

Item dimensionality:

Exploratory Factor Analysis (EFA)

Scale dimensionality

- Generally, when we aim to create a composite score, we want it to reflect one single concept or attribute
  - For example, we would not add trust + usability to form one score

- However, if we had a concept with multiple dimensions, we could report a composite score for the entire scale, or for each dimension
  - For example, if we had a workload scale reflecting both mental and physical dimensions, we could report:
    - » **one** workload score OR
    - » a score for **physical workload** and a score for **mental workload**

- If a scale has multiple dimensions, the correlation between the dimensions should be evaluated when deciding whether to “roll up” the dimensions into one composite score

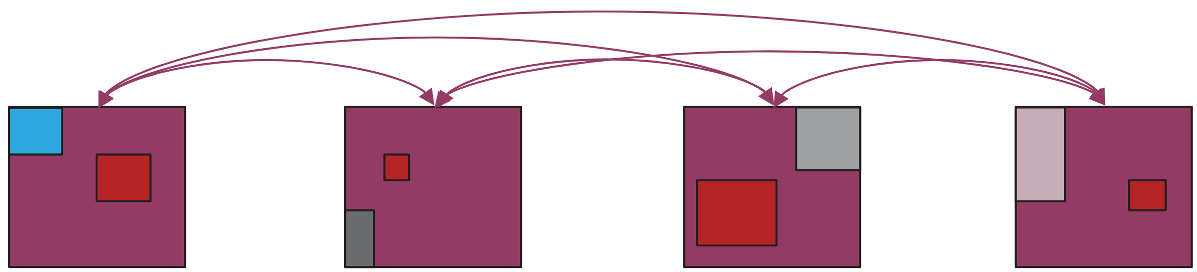
Reasoning of EFA

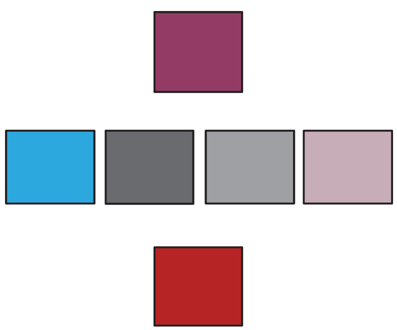
- If two items are highly correlated, it’s because they are both dependent on an underlying factor

- We seek to identify the **number** and the **nature** of factors that produce the correlation we see in our items.

- EFA can be used to identify dimensions within the concept we are measuring

Suppose I have four correlated items…





Common variance

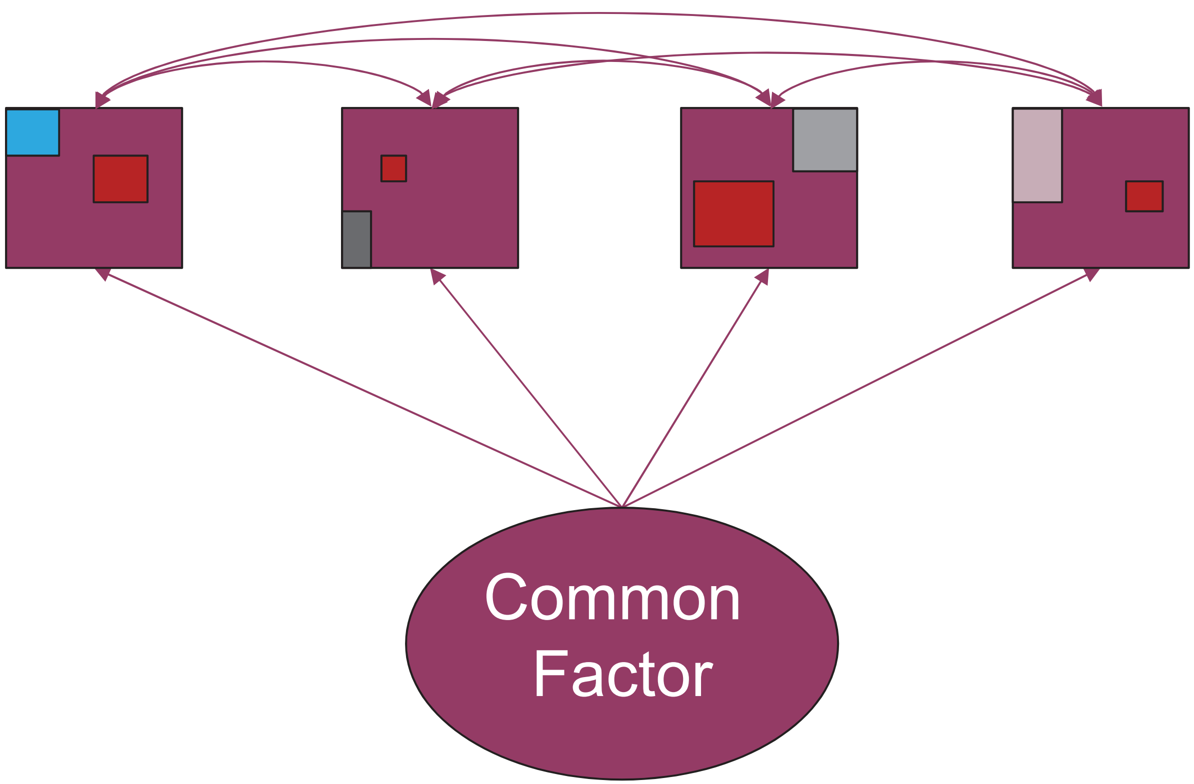
Specific variance

Error variance

In words: each survey item is composed of some amount of **common** variance (related to the concept we are measuring), some amount of variance **unique** to that item, and some amount of variance that doesn't have to do with the item at all (**error**).

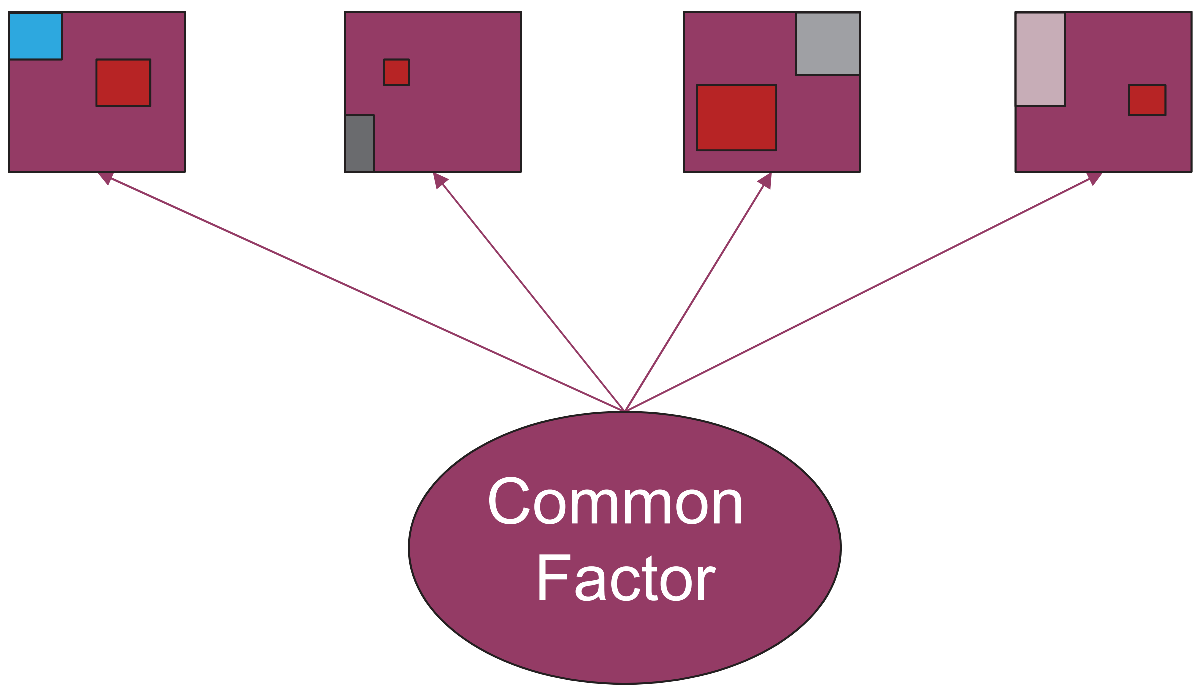
Exploratory Factor Analysis

Key idea – the correlations are explained by the fact that the variables have a common underlying "factor"



EFA, continued

And we are no longer interested in the remaining correlations among the variables



We can conceptually partition the variance of a single item

Observed Variance of Item

Common Variance

Specific Variance

Error Variance

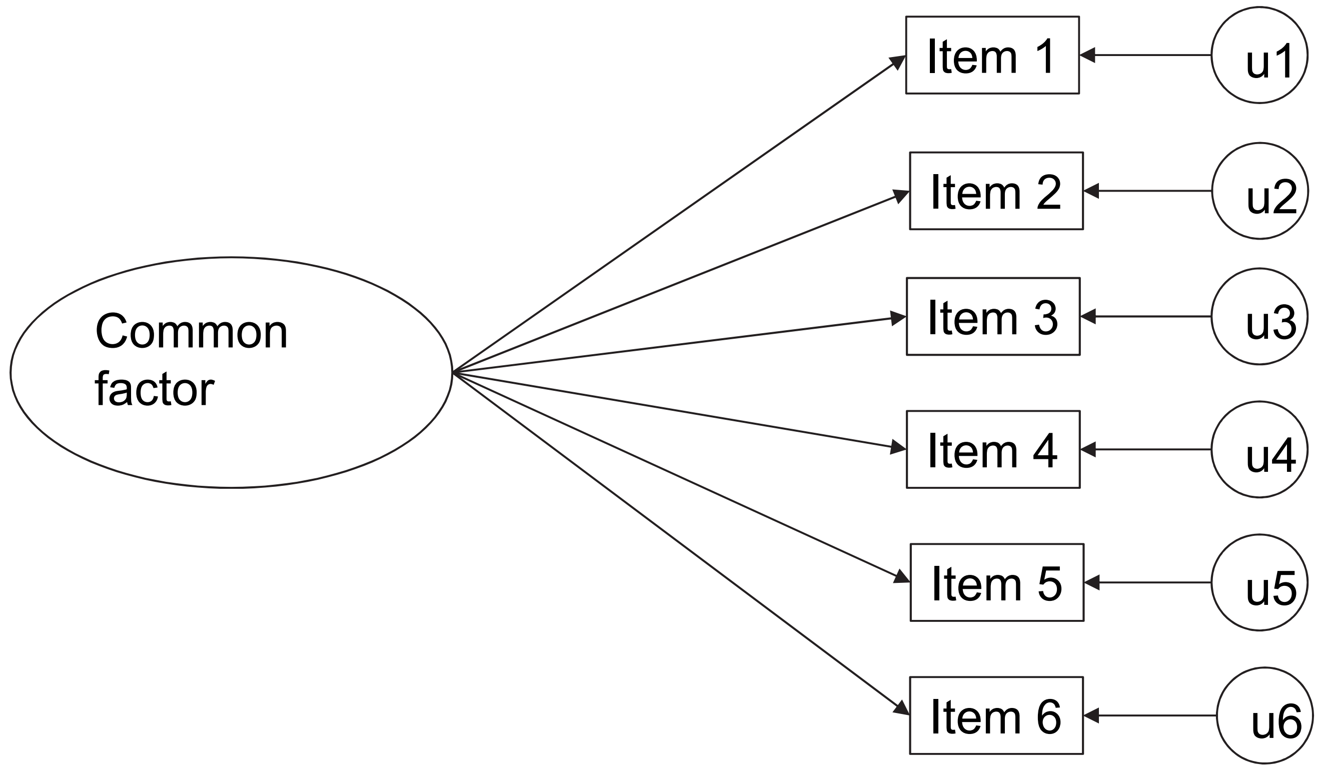
Unique Variance

Reliable Variance

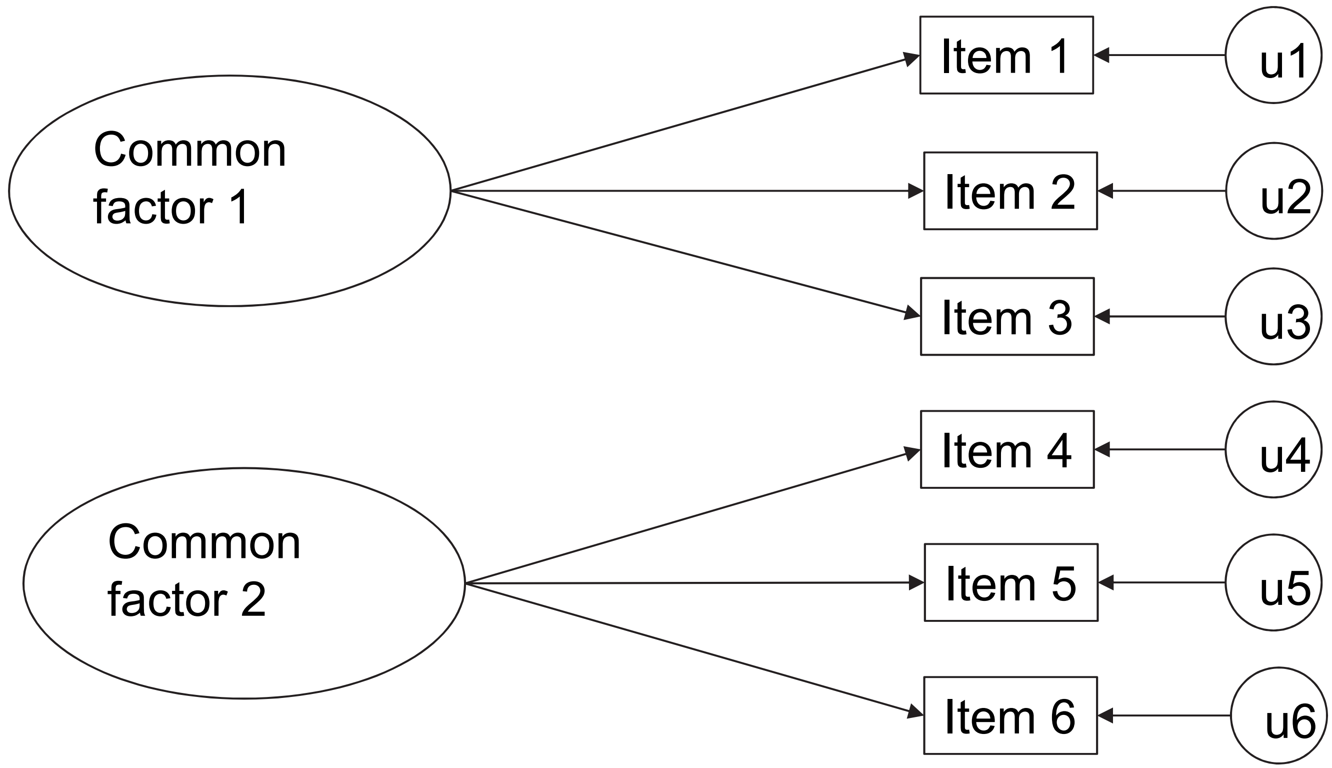
Reliability: (common + specific)/observed variance

Suppose we have six items. We want to know if these six items represent one or two dimensions of our construct.

Does one common factor adequately explain the correlations we see among the items?



Or do two common factors better explain the item intercorrelations?



Exploratory factor analysis provides an answer to this question.

Mathematical expression

- $x_{ij} - \mu_j = \lambda_{j1}z_{i1} + \lambda_{j2}z_{i2} + \ \ldots + \lambda_{jm}z_{im} + 1u_{ij}$

- x_{ij} is the item score for person i on variable j
- μ_j is the mean of variable j
- λ_{jk} is the factor loading of item j on factor k
- z_{ik} is the common factor score for person i on factor k
- u_{ij} is the factor score for person i on unique factor j

In a sentence: an individual's score on an item is a linear combination of individual scores on a common factor plus the effect of a unique factor.

That expression looks familiar… we can think of it like a multiple regression

$$x_{ij} - \mu_j = \lambda_{j1} z_{i1} + \lambda_{j2} z_{i2} + \ \ldots + \lambda_{jm} z_{im} + 1 u_{ij}$$

x_{ij} items are like dependent variables (means subtracted)

z_{ik} factor variables are like independent variables

λ_{jk} factor loadings are like regression weights

u_{ij} are like error terms

Except our independent variables (our latent factors) are unobserved, and must be estimated

For our six notional items with two factors, we could represent those items as follows:

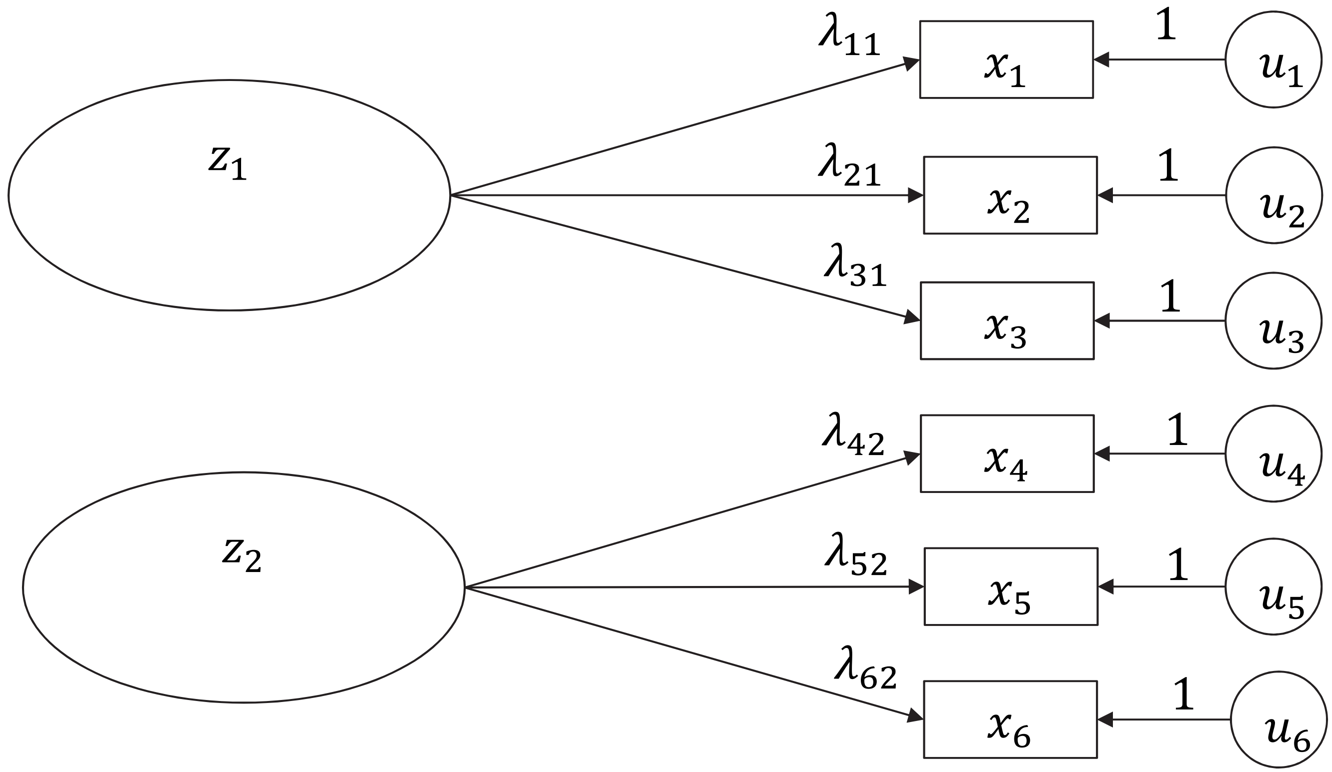
$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} + \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \\ \lambda_{61} & \lambda_{62} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix}$$

If we wanted to express the equation for item 2:

$$\begin{bmatrix} x_2 \\ \\ \\ \\ \\ \end{bmatrix} = \begin{bmatrix} \mu_2 \\ \\ \\ \\ \\ \end{bmatrix} + \begin{bmatrix} \lambda_{21} & \lambda_{22} \\ & \\ & \\ & \\ & \\ & \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} u_2 \\ \\ \\ \\ \\ \end{bmatrix}$$

$$x_2 = \mu_2 + \lambda_{21} z_1 + \lambda_{22} z_2 + u_2$$

Placed back on our path diagram…



$$x_{ij} - \mu_j = \lambda_{j1} z_{i1} + \lambda_{j2} z_{i2} + \ \ldots + \lambda_{jm} z_{im} + 1 u_{ij}$$

From my six items, how do I decide on the number of factors?

- There are multiple tools available to you when deciding the number of factors to retain.

- The scree plot is a popular visual tool. It makes use of eigenvalues.

- Other rules:
  - # Eigenvalues > 1
  - Goodness of fit measures
  - Information criteria
  - Interpretability of factors

The role of eigenvalues and eigenvectors in understanding EFA

Suppose we have a correlation matrix, $\boldsymbol{S}$

We can represent this correlation matrix in terms of eigenvalues and eigenvectors, where $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D_l}\boldsymbol{U}'$

$\boldsymbol{D_l}$ is a diagonal matrix containing eigenvalues

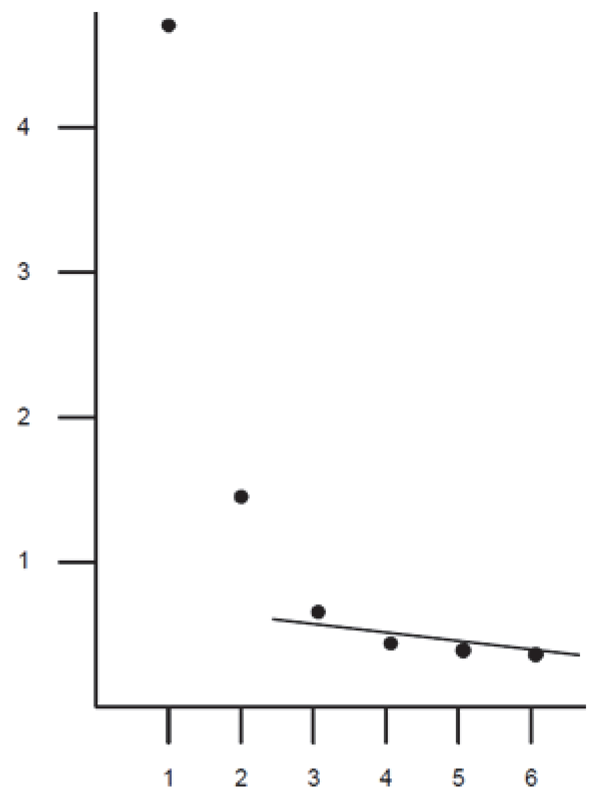
$\boldsymbol{U}$ is a square matrix containing eigenvectors

The number of eigenvalues is equal to the number of items

These eigenvalues represent the amount of variance accounted for by each factor

A scree plot helps you decide the number of factors to retain using eigenvalues

- Look for a distinct drop – specifically, look for the last large drop in the series.



We would retain two factors, as the first two factors explain most of the variance in our items.

Putting this together:
A notional example

Suppose we create a new scale with the following items:

- The system is reliable.
- The system is dependable.
- I believe the output of the system.
- The system behavior is predictable.
- The system processes information quickly.
- The system is timely when delivering information.
- I can finish my task on time using the system.
- I rarely experience system delays.
- The system is trustworthy and efficient.

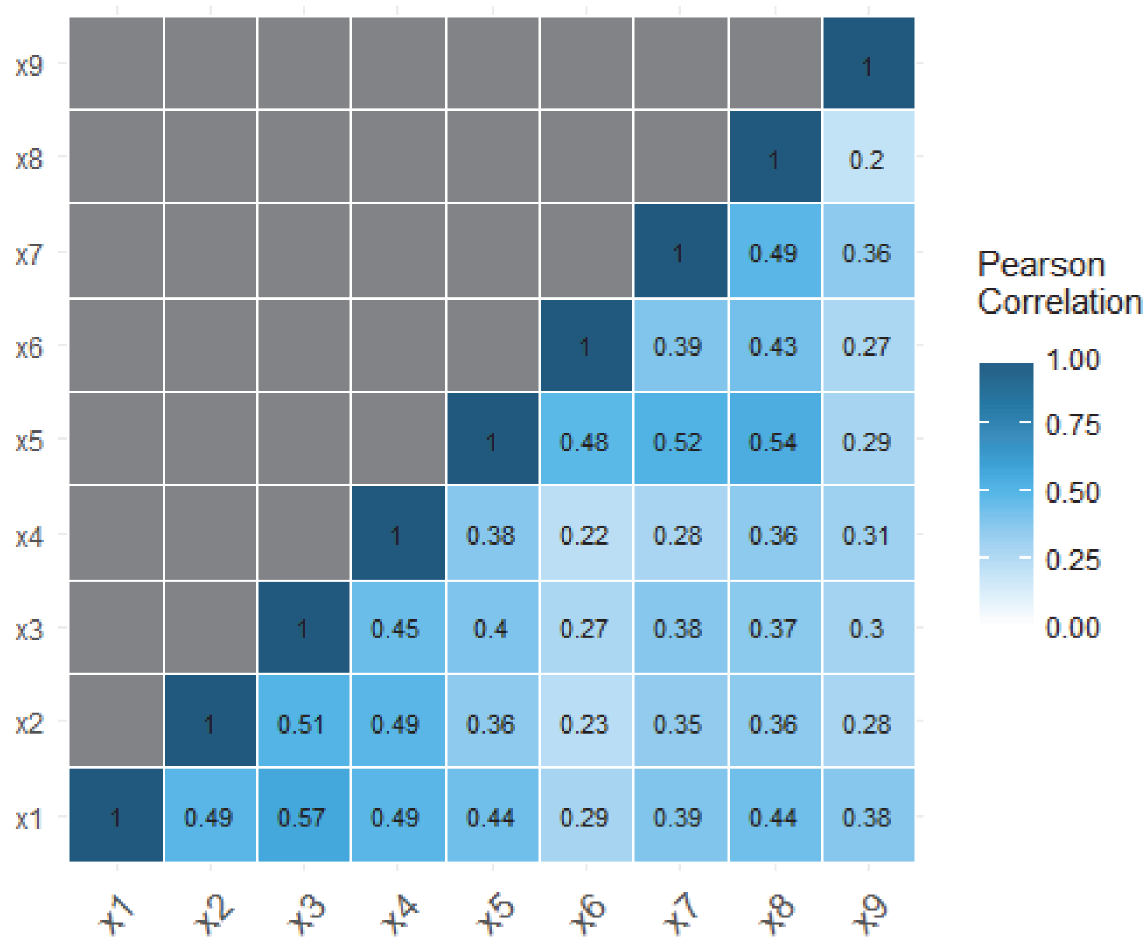
Items are rated on a 1 – 7 Likert scale from “Strongly Disagree” to “Strongly Agree.”

We conduct survey pre-testing and administer the survey to a number of target operators (N = 100) after using the system.

We pull our data into our favorite software program of choice [let’s say JMP].

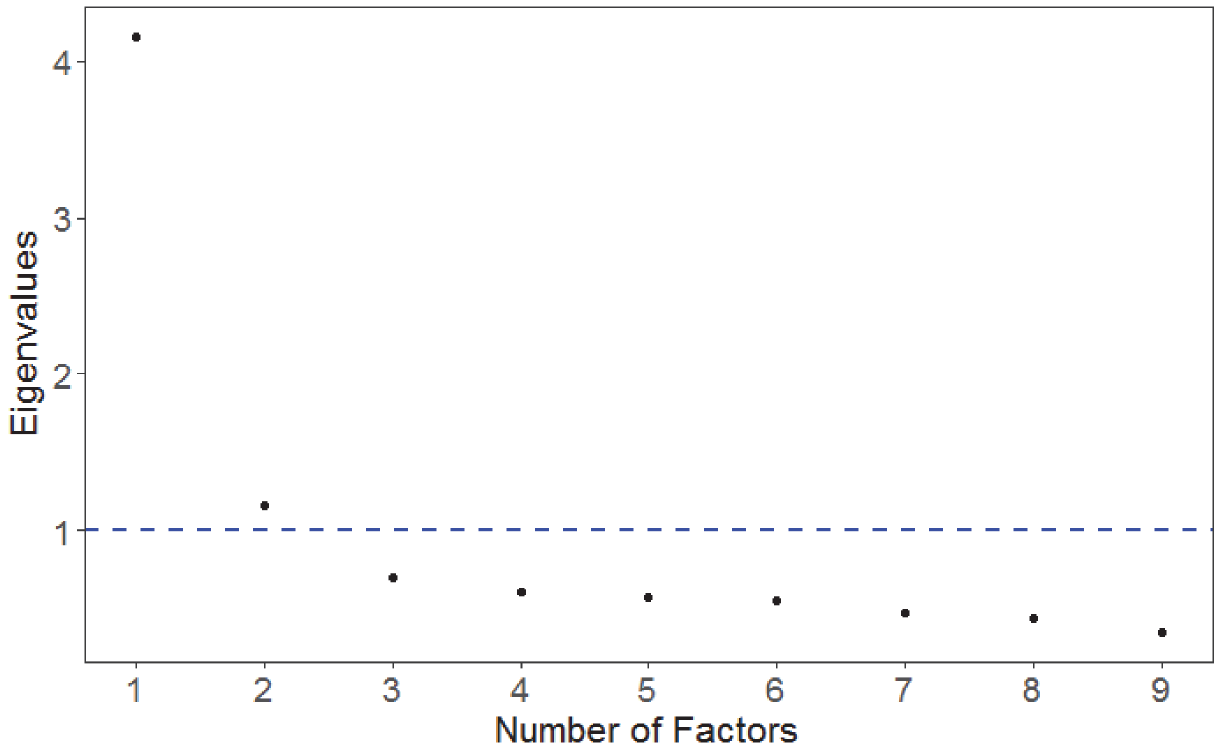
We start by examining a correlation matrix, and then a scree plot.

A correlation matrix shows us the strength of linear relationship among our variables



Deciding the number of factors

Next, we examine a scree plot of the eigenvalues of our item correlation matrix. How many factors should we retain?



We would retain two factors

When doing EFA among related constructs, oblique factor rotation is preferred

Oblique Rotation

Orthogonal Rotation

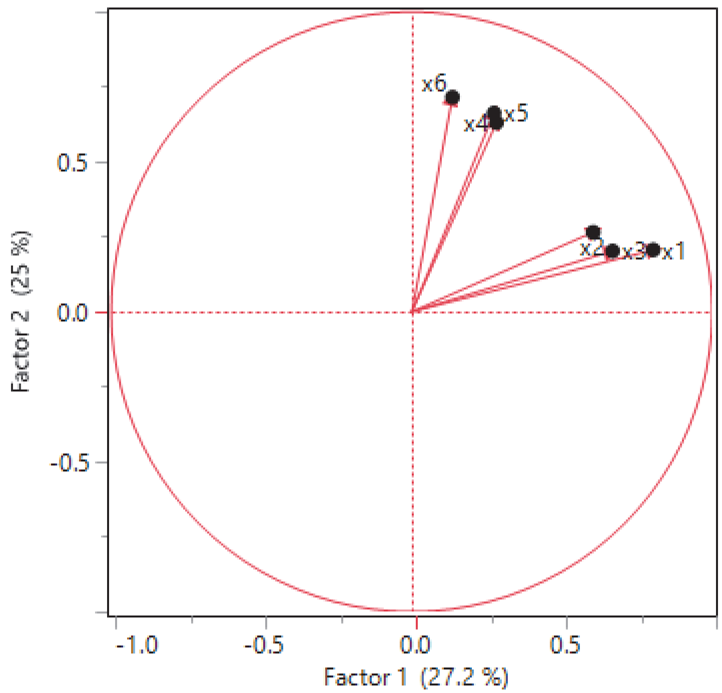
Oblique factor rotation means that your factors are allowed to correlate.

Forcing orthogonality (no correlation) will result in poorly defined factors.

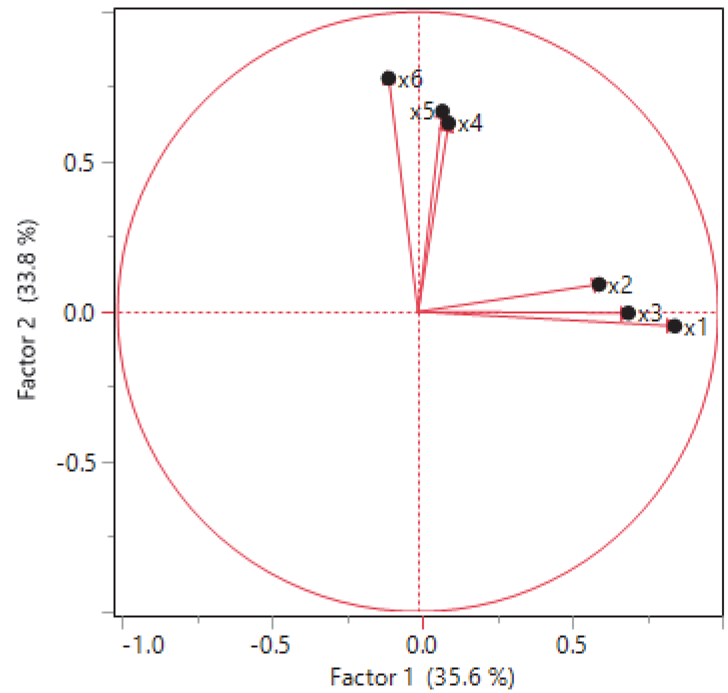
Visualizing factor rotation

If we force the factors to be uncorrelated, our items will load on both factors.

We can see here that these two factors are correlated, as the angle between them is less than 90 degrees.



By allowing the factors to correlate, the items load cleanly on each factor, and each factor explains more variance in the items.



Results

Next, we conduct an EFA, telling the software program to extract two factors. We obtain the following results:

Factor loading matrix:

	Factor 1	Factor 2
x1	0.75	0.02
x2	0.70	-0.02
x3	0.66	0.08
x4	0.74	-0.09
x5	0.03	0.76
x6	-0.10	0.70
x7	0.07	0.67
x8	0.09	0.64
x9	0.36	0.15

Factor correlation matrix:

	Factor 1	Factor 2
Factor 1	1.00	0.67
Factor 2	0.67	1.00

**Helpful hint: remember that factor loadings are like regression coefficients.*

Results, continued

We see evidence of two related, but distinct factors with “simple structure.”

Factor loading matrix:

	Factor 1	Factor 2
x1	0.75	0.02
x2	0.70	-0.02
x3	0.66	0.08
x4	0.74	-0.09
x5	0.03	0.76
x6	-0.10	0.70
x7	0.07	0.67
x8	0.09	0.64
x9	0.36	0.15

Factor correlation matrix:

	Factor 1	Factor 2
Factor 1	1.00	0.67
Factor 2	0.67	1.00

**Helpful hint: remember that factor loadings are like regression coefficients.*

Results, continued

Factor loading matrix:

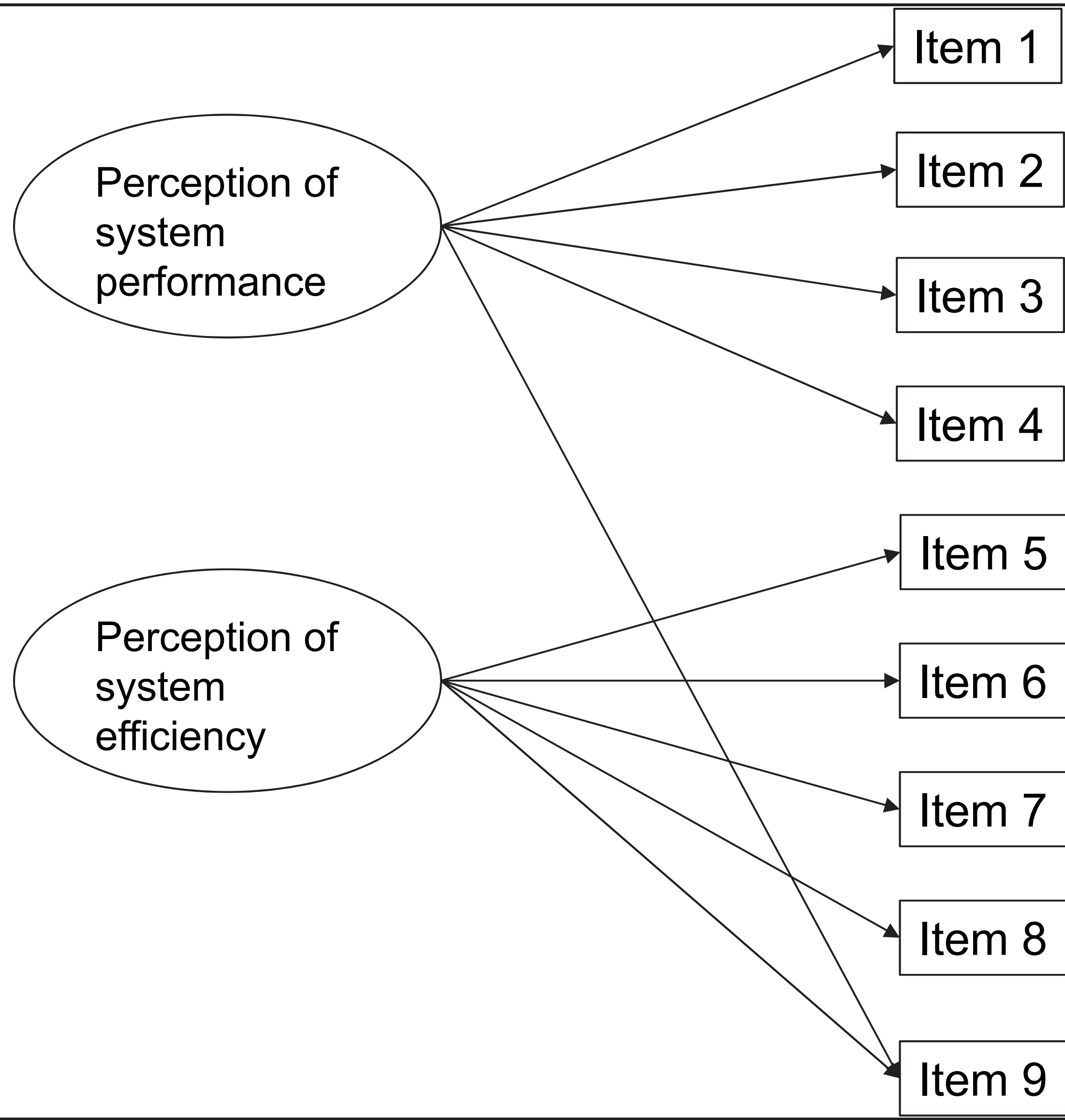
	Factor 1	Factor 2
X1: The system is reliable.	0.75	0.02
X2: The system is dependable.	0.70	-0.02
X3: I believe the output of the system.	0.66	0.08
X4: The system behavior is predictable.	0.74	-0.09
X5: The system processes information quickly.	0.03	0.76
X6: The system is timely when delivering information.	-0.10	0.70
X7: I can finish my task on time using the system.	0.07	0.67
X8: I rarely experience system delays.	0.09	0.64
X9: The system is trustworthy and efficient.	0.36	0.15

**Helpful hint: remember that factor loadings are like regression coefficients.*

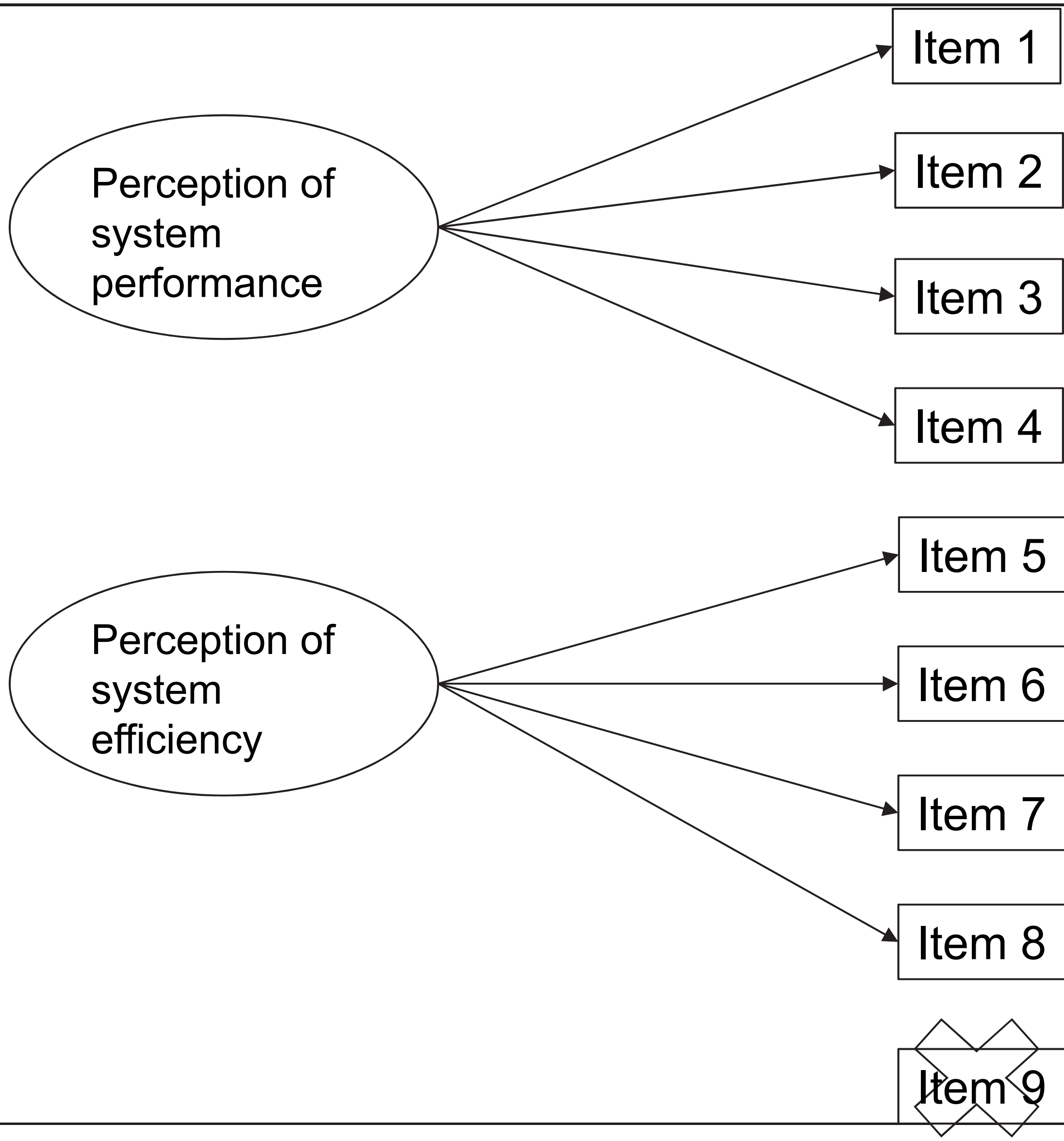
In this scale, we see evidence that two related latent factors underlie our nine items.

We would drop item 9, as it does not “load cleanly” on either factor.

Summarized graphically: Our EFA gave us these results…



And we move forward with this model

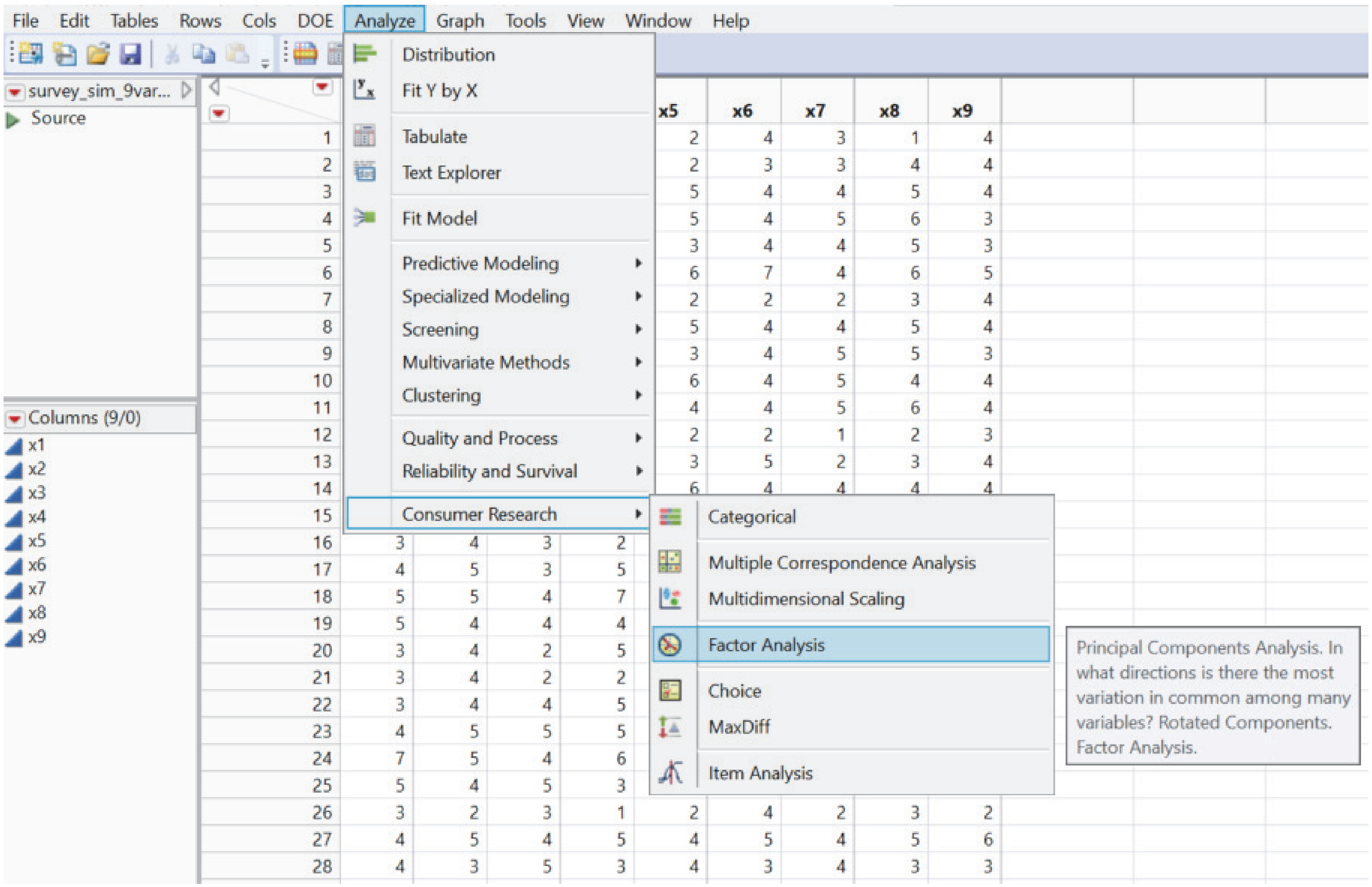


We have now investigated the dimensionality of our scale.

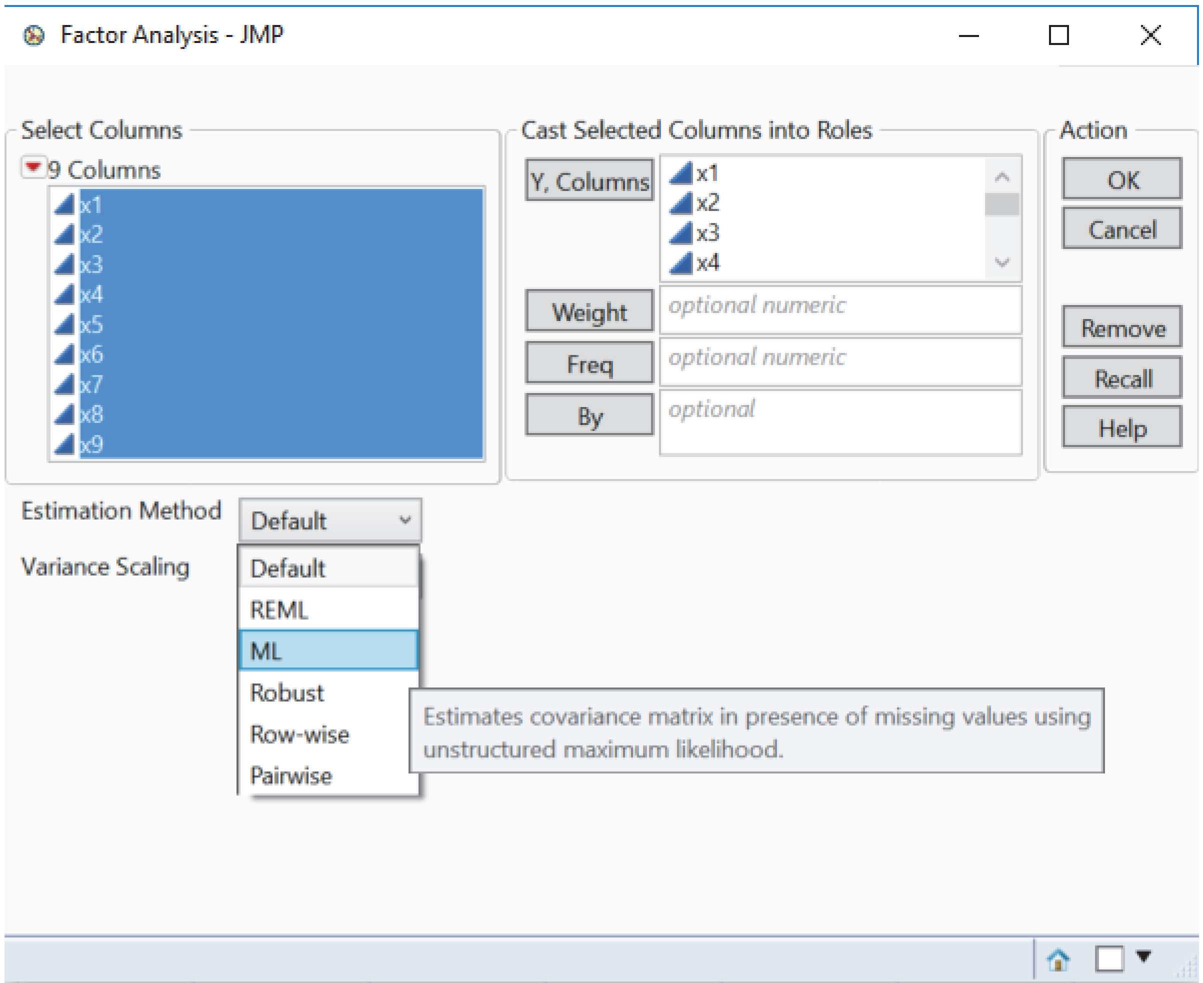
We can report two scores from our scale: perception of system performance and perception of system efficiency.

IDA

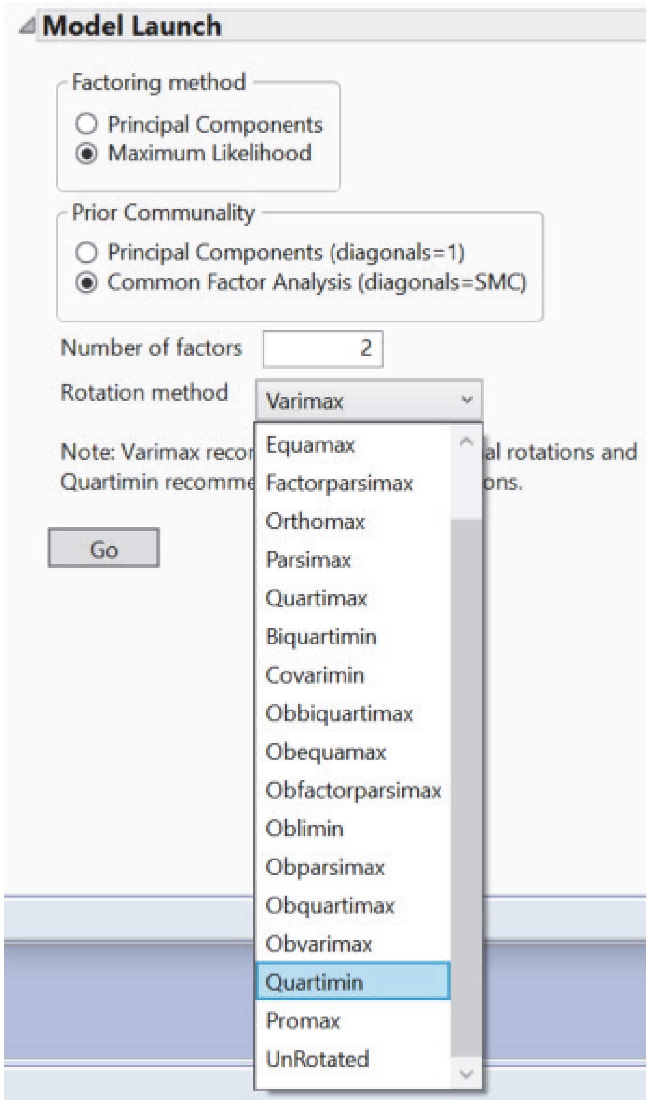
Conducting an EFA in JMP

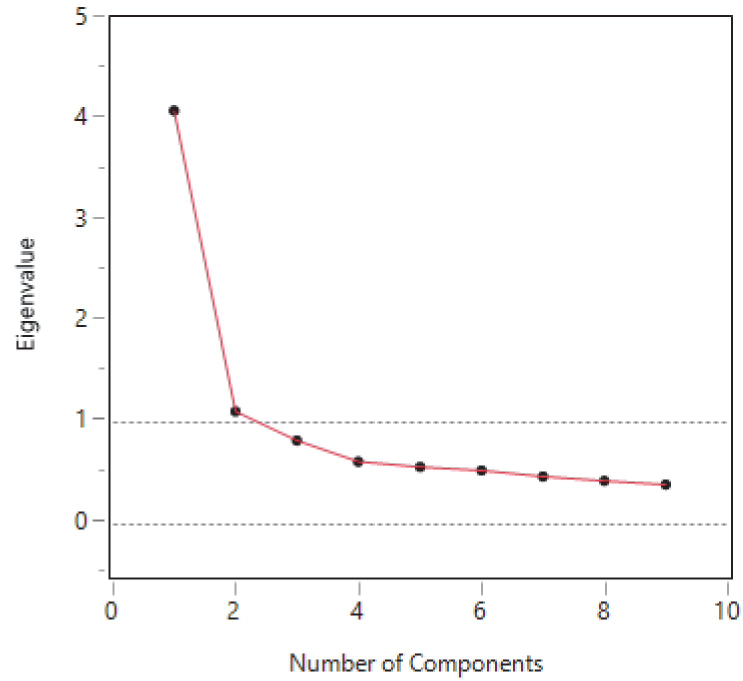


Conducting an EFA in JMP



Conducting an EFA in JMP





Rotated Factor Loading

	Factor 1	Factor 2
x3	0.727910	-0.002712
x1	0.719398	0.058166
x2	0.709802	-0.033407
x4	0.680964	-0.030271
x9	0.341252	0.155652
x5	0.067873	0.717988
x6	-0.099162	0.685834
x8	0.102164	0.631573
x7	0.095435	0.616541

But what if we don't seek to *explore* the factor structure of our scale?

What if we've already formed a hypotheses about the structure of our scale, and we want to test that *confirmatory* hypothesis?

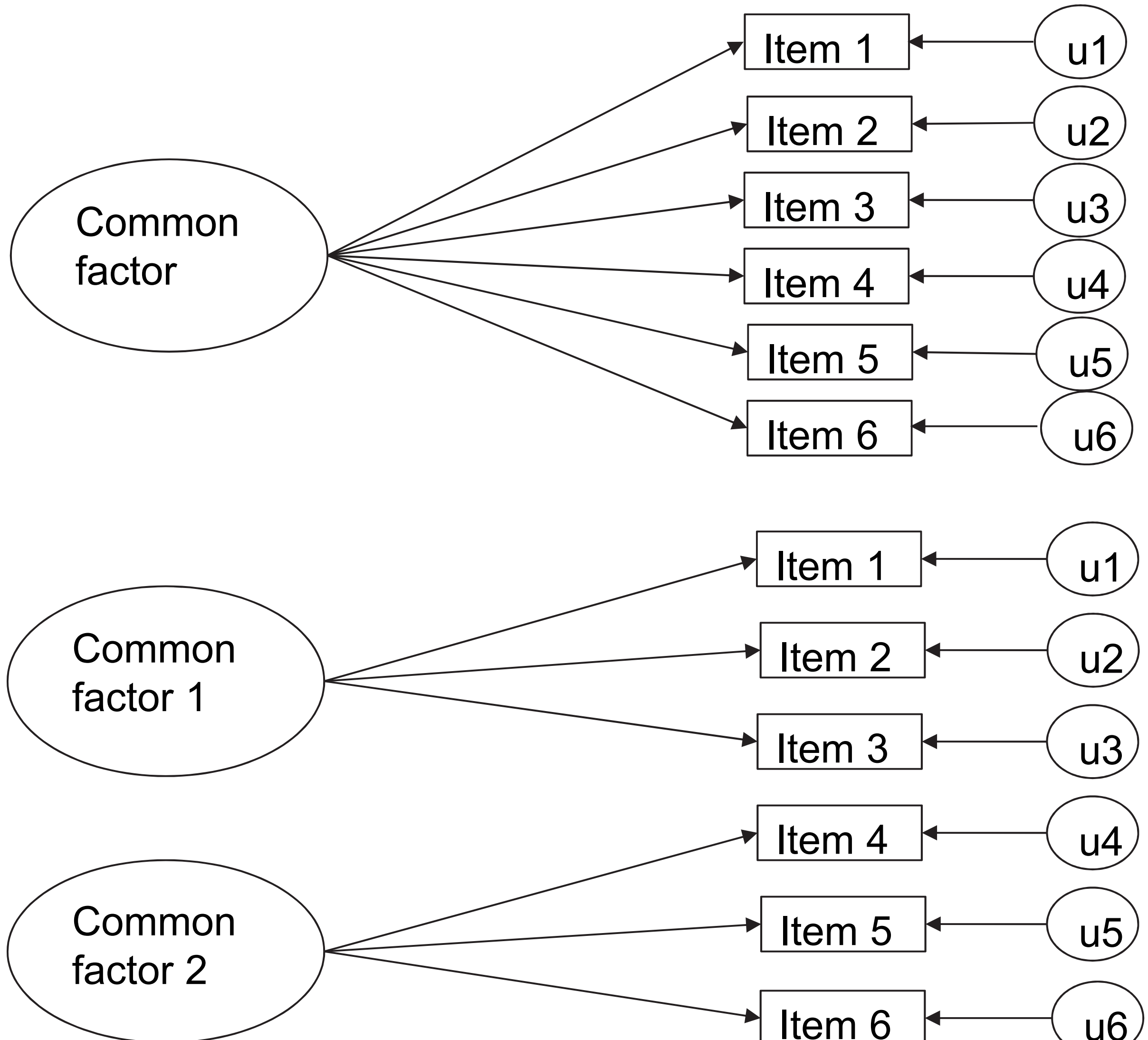
Confirmatory Factor Analysis (CFA)

- In EFA, we seek to discover the number and the nature of factors that underlie the correlation matrix we see

- In CFA, we seek to test an explicit hypothesis, sometimes competing hypotheses, about the number and nature of factors that underlie the item correlations

- CFA is a powerful modeling technique that requires *a priori* hypotheses

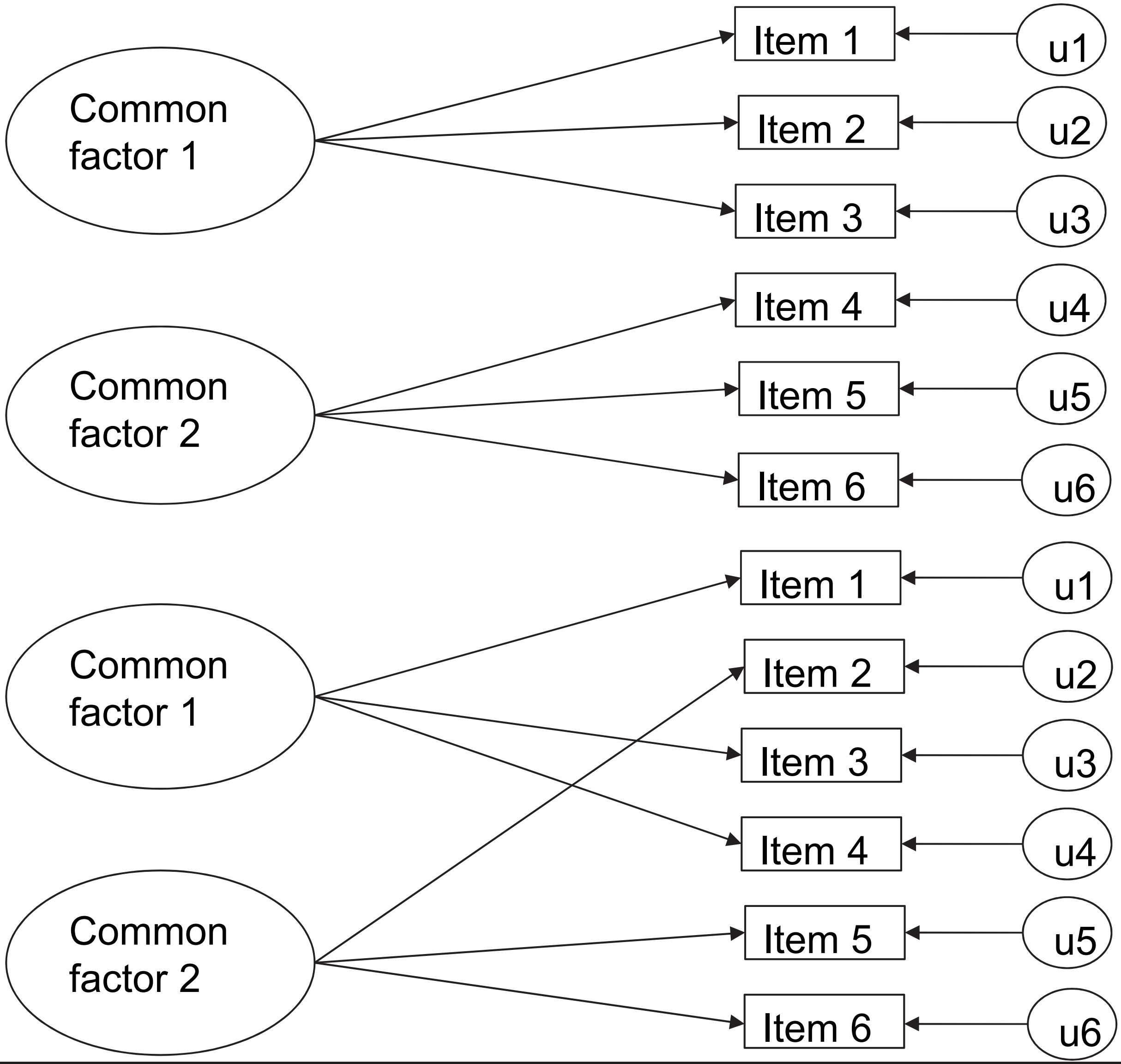
Competing models:
One factor versus two factors



CFA can be used to decide between a one- or multi-factor structure

Competing models: Different two-factor configurations

CFA can also be used to decide between competing *k*-factor structures



Confirmatory Factor Analysis

- The results of confirmatory factor analysis provide evidence that allows us to decide between a competing number of factors or competing factor structure to represent our constructs

- A benefit of CFA is that you (the researcher) are less likely to capitalize on sample-specific idiosyncrasies (e.g., chance) when investigating factor structure

- One downside of CFA is that it is less readily available in software such as JMP. Instead, it involves using syntax available in a user-contributed R package, *lavaan,* or in a standalone program such as Mplus.

Key takeaways

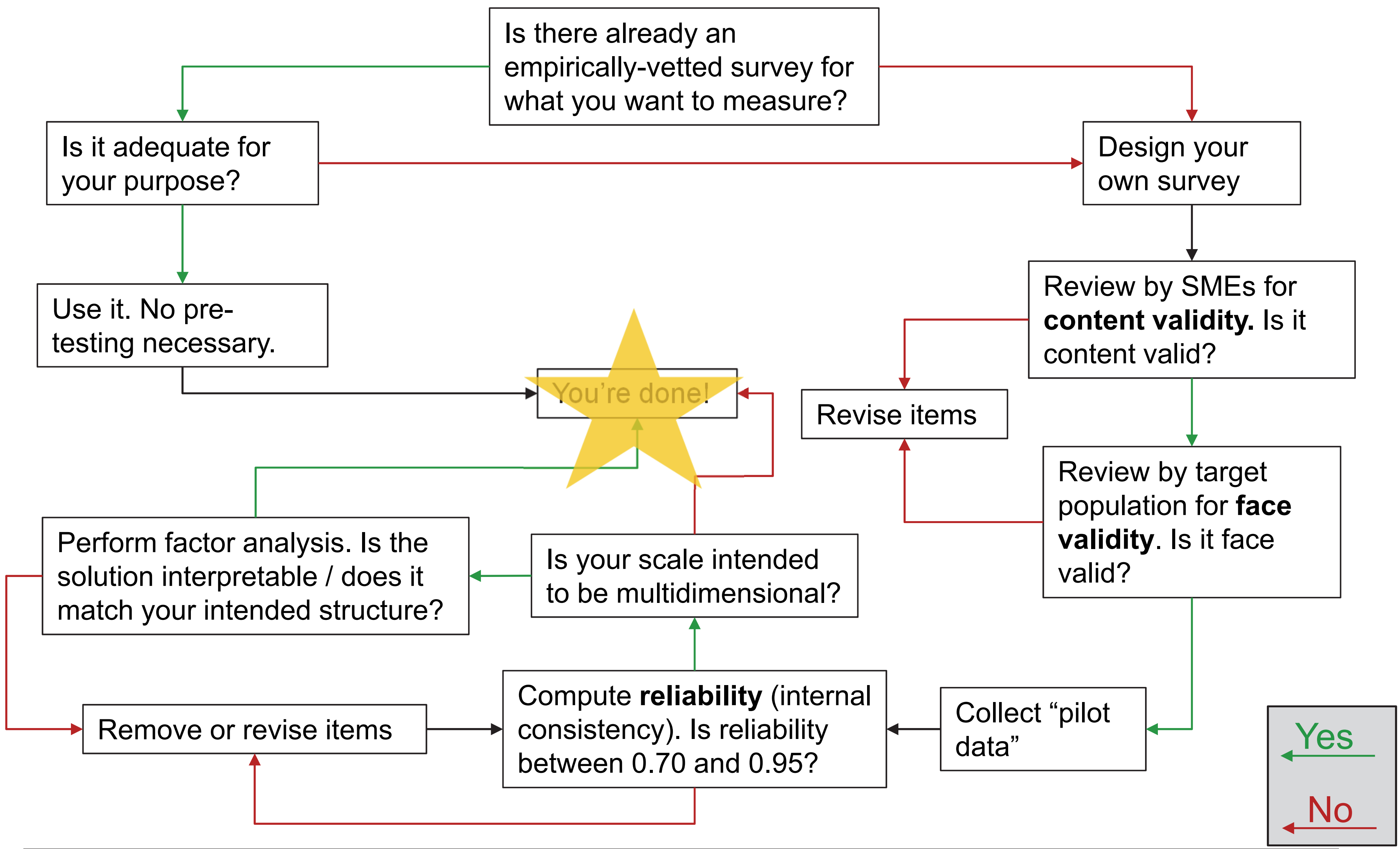
Summary points

- Face validity and content validity should be established before pilot data are collected. Criterion validity and internal consistency should be established after the collection of pilot data.

- Cronbach's alpha is a measure of how well our items hang together, and can be used to justify a roll-up calculation of a set of items. However, it is not a test of dimensionality.

- EFA should be used when we want to investigate the dimensionality of our scale, and when we want to validate a unidimensional or multidimensional scale. CFA should be used when we want to do so with explicit hypotheses.

Roadmap to survey success



REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
31-12-2018

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
January 2018 - December 2018

4. TITLE AND SUBTITLE
Vetting Custom Scales - Understanding Reliability, Validity, and Dimensionality

5a. CONTRACT NUMBER
HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Lane, Stephanie, OED

5d. PROJECT NUMBER
BD-9-2299

5e. TASK NUMBER
2299(90)

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER
D-9168-NS
H 2018-000279

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Director, Operational Test and Evaluation
The Pentagon
1700 Defense
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)
DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
This draft has not been approved by the sponsor for distribution and release. Reproduction or use of this material is not authorized without prior permission from the responsible IDA Division Director.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
For these situations in which an empirically vetted scale does not exist or is not suitable, a custom scale may be created. This document presents a comprehensive process for establishing the defensible use of a custom scale. At the highest level, this process encompasses (1) establishing validity of the scale, (2) establishing reliability of the scale, and (3) assessing dimensionality, whether intended or unintended, of the scale. First, the concept of validity is described, including how validity may be established using operators and subject matter experts. The concept of scale reliability is described, with guidelines for computing, interpreting, and using results to inform potential modifications to a custom scale. Next, a method for investigating the dimensionality of a scale, exploratory factor analysis, is described, along with a walkthrough of software implementation and results. Finally, confirmatory factor analysis, a technique for testing a priori hypotheses about dimensionality, is presented.

15. SUBJECT TERMS
Dimension Reduction; Reliability; Scale Validity; Survey

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
Unclassified	Unclassified	Unclassified	Unlimited	103	Dr. Heather Wojton, Project Leader, OED
					19b. TELEPHONE NUMBER *(Include area code)* 703.845.6811