Human-Computer Interaction

Measurement I: Principles, Subjective Measures, & Scale Construction

Professor Bilge Mutlu

Today's Agenda

» Measurement Principles

» Subjective Measures

» Scale Construction

Measurement Principles

What do we measure when we measure?

Definition: Measurement is the assignment of a number to a characteristic of an object or event, which can be compared with other objects or events.[1]

But it's notjust numbers! We can measure:

Quantitative measurements describe the degree of an attribute, e.g., an under-three-hour marathon runner, someone who scores 1600 in SAT

Qualitative measurements describe subjective observations, e.g., "the first customer was a teenage boy"

[1] Wikipedia

How do we represent measurements?

We measure **variables** — things that can change
→ *e.g.*, preference, accuracy, age

Each variable has **attributes** — possible values or categories
→ *e.g.*, high / low performance, 18–25 / 26–35 / 36–45

What are different types of variables?

From least to most precise: *Nominal* → *Ordinal* → *Interval* → *Ratio*

Nominal: Categories or labels
Example: novice / expert, male / female
→ Can distinguish, but not rank

Ordinal: Ordered categories
Example: very satisfied → very unsatisfied
→ Can rank, but distances unequal

Interval: Equal intervals, no true zero
Example: temperature, satisfaction 1–7
→ Can compare differences

Ratio: Equal intervals, real zero
Example: time, weight, task completion count
→ Can compare ratios (“twice as fast”)

What are different kinds of measurements we can take?

1. **Objective**: Measurement from participants against an objective standard, e.g., performance in a test, typing speed

2. **Behavioral**: Measurement of the actions and behaviors of participants, E.g., how much eye-contact participants maintain with a robot

3. **Subjective**: Measurement of self-report data on subjective evaluations, e.g., preferences, personality 👈 **our focus today**

4. **Physiological**: Measurements taken directly from participants' bodies, e.g., body temperature, GSR, EEG, EMG, fMRI

What makes measurements good?

1. Validity
2. Reliability
3. Quality

What is validity?

Definition: Validity is the extent to which a concept, conclusion, or measurement is well-founded and likely corresponds accurately to the real world.[2]

In other words, are we measuring what we want to measure?

[2] Wikipedia)

What is an example validity problem?

Consider wanting to measure **aggression in children**

We can measure the amount of time children play with: *aggressive toys* (guns, swords, tanks) vs. *non-aggressive toys* (trucks, tools, dolls)

What are threats to validity?

» Children might be playing with toys that they are more familiar with, e.g., they see guns and tanks on TV all the time

» Children can also play with trucks and dolls in aggressive ways

What are different forms of validity?

Type	Definition	Example / Key Idea
Face validity	Does the measure *appear* valid based on logic and judgment (not statistics)?	A motivation scale that *looks* like it measures motivation.
Construct validity	Does the measure capture the intended *conceptual construct*?	·**Convergent:** correlates with related constructs (e.g., *extroversion ↔ sociability*) · **Discriminant:** not correlated with unrelated constructs (e.g., *intellect ≠ control*)
Empirical (criterion) validity	Does the measure relate to *established* measures or predict outcomes?	·**Concurrent:** correlates with other measures taken at the *same time* · **Predictive:** predicts future measures or outcomes
Content validity	Does the measure represent a *comprehensive sample* of the construct's aspects?	GRE verbal test covers vocabulary but misses grammar and comprehension.
Ecological validity	Do the results *generalize* to real-world contexts?	Detecting dots on a screen may not translate to detecting cars in traffic.

What is a construct?

Definition: A psychological construct is a label for a cluster or domain of covarying behaviours.[3]

What is a scale, or a psychometric scale?

Definition: A single item measured with a variety of response formats (e.g., semantic differential, multiple choice, checklist) or a collection of items with similar formats.[4]

What happens when we have low or high construct validity?

A high construct validity means that the measurement sufficiently captures and covers the construct.

---

[3] Brittanica

[4] APA

A measure of ***introversion*** $(\alpha = .85)$[5]

Positive keyed

» Don't like to draw attention to myself.

» Keep in the background.

» Dislike being the center of attention.

» Don't talk a lot.

Negative keyed

» Don't mind being the center of attention

» Take charge.

» Want to be in charge.

» Am the life of the party.

» Can talk others into doing things.

» Seek to influence others.

---

[5] 5 IPIP scale of introversion ↩ ↩, IPIP scales

What is reliability?

Definition: Reliability in statistics and psychometrics defines the consistency of a measure across repeated measurements and judgments.

E.g., more robot eye contact leads to better information recall; could we replicate this result with a second set of subjects or with the same subject another time?

High reliability indicates that the measure produces similar results under consistent conditions.

Reliability is decreased by error.

What are different forms of realiability?

Type	Definition	Key Methods / Notes
Test–retest reliability	Repeating the same measurement with the same population at a later time.	• Simple to administer • Risk of memory effects or true change over time
Alternative-form reliability	Administering a similar but not identical measure to the same population.	• Reduces memory bias • Lower reliability if new items differ too much
Split-half reliability	Splitting one test into two halves and correlating results.	• Done in a single session • Avoids repeat testing issues

Broader Reliability Types	Definition	Common Measures
Internal reliability[6]	Checks whether multiple items measuring the same construct produce similar results.	• *Inter-item correlation* (average item relationships) • *Split-half correlation* (random halves) • *Cronbach's α* (overall internal consistency, α > .70 desirable)
Inter-coder reliability	Assesses agreement among raters observing the same phenomenon.	• Percent agreement • Cohen's κ, Fisher's κ, Krippendorff's α

[6] Wikipedia

How do we calculate reliability?

We can't—we can only *estimate* it using statistical methods.

$$R = \frac{v_{measurement}}{v_{true} + v_{error}}, R \in [0, 1]$$

Where $v_{measurement}$, v_{true}, and v_{error} are the variances on the measured, true and error scores, respectively. v_{true} can never be measured, so it's estimated.

Pro Tip: A reliability of .70 or higher is acceptable.

What does data quality mean?

Data quality is affected by **measurement error** (or observational error).

Definition: The difference between the measurement (what is recorded) and the true quantity of the variable, i.e., distortions that cause the observed measurement to be different from the true quantities.

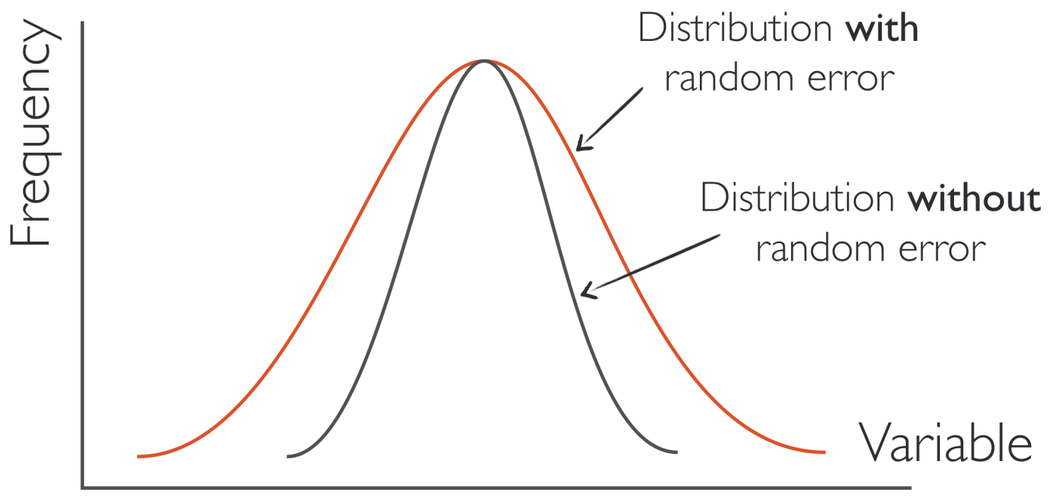
$$X = T + e_r + e_s$$

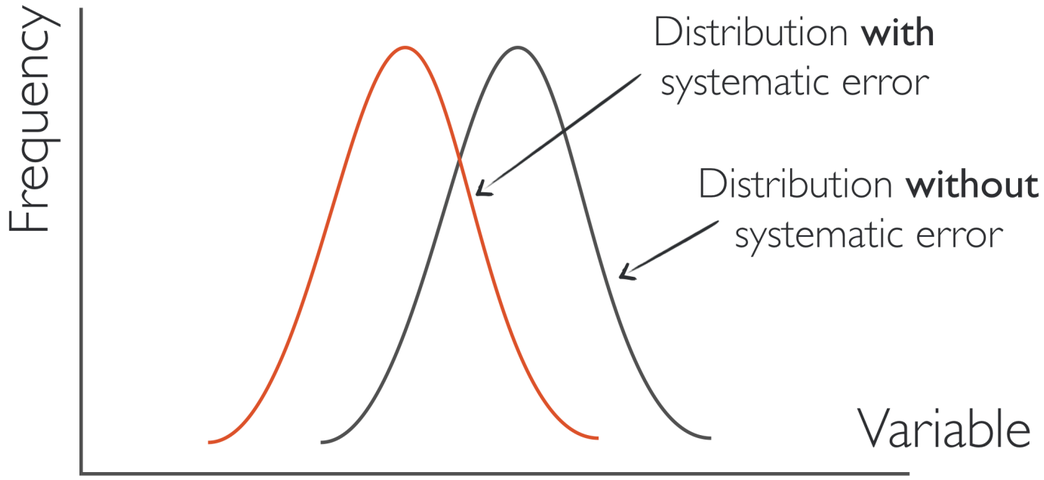
What are different types of error?

Type	Definition	Causes / Effects
Random Error (Also called "noise")	Inconsistent variation in repeated measurements of the same attribute.[6]	• Inherent variability in participants or conditions (e.g., prior experience) • Affects **variance**, not the mean
Systematic Error (Also called "bias")	Consistent inaccuracy introduced by flaws in the measurement process.[6]	• External factors such as equipment delay, coder bias, or calibration issues • Affects the **mean**, not variance

[6] Wikipedia

Random vs. systematic error





How do we increase measurement quality?

Method	Definition / Purpose
Piloting experimental instruments	Conduct small-scale test runs of surveys or experiments to detect flaws, ambiguities, or technical issues before full deployment.
Testing reliability of coders, retraining	Evaluate agreement among raters and retrain when inconsistencies arise to maintain coding accuracy and consistency.
Repeating manual data entry	Enter data twice or verify through cross-checking to identify and correct transcription or input errors.
Measuring error using statistical methods	Quantify measurement error or variance through reliability coefficients or error modeling to assess data quality.
Triangulation	Use multiple methods, measures, or data sources to cross-validate findings and strengthen the credibility of results.

Subjective Measures

What are subjective measures?

Definition: Capture people's *perceptions*, *thoughts*, *feelings*, and *preferences* through self-report instruments.

Common instruments

» **Questionnaires** → standardized responses (e.g., rating satisfaction)

» **Interviews** → open-ended responses (e.g., exploring experiences)

These can be administered by the *researcher* (interview) or *respondent* (survey form).

A **survey** is a quantitative method using questionnaires or interviews to collect and analyze data from a population.

What types of questions are asked in subjective measures?

Open-ended questions

» Invite rich, unstructured responses

» Useful for exploring *why* or *how* people think or act

Example: "What barriers did you face in using the Banjee software?"

Closed-ended questions

» Provide structured response options for statistical analysis

» Useful for comparing *what* or *how much*

Example: "Rate your experience using Veggieworld.com"[7]
(Frustrating 1 — 2 — 3 — 4 — 5 — 6 — 7 Satisfying)

[7] Lazar et al., 2017, Chapter 5 - Surveys

What are different standardized response instruments?

Structured formats for measuring subjective evaluations:

1. **Likert scales** — agreement with statements
2. **Rating scales** — numeric ratings of intensity
3. **Semantic differential scales** — bipolar adjective pairs

What are some examples?

Likert scale: "I find this robot trustworthy."

Strongly disagree — Disagree — Neutral — Agree — Strongly agree

Rating scale: "Rate your satisfaction."

(Frustrating) 1 — 2 — 3 — 4 — 5 — 6 — 7 (Satisfying)

Semantic differential scale: "Using this system is…"

Easy⊙⊙⊙⊙⊙Difficult

What are scales (or metrics)?

A **scale** (or **metric**) measures a psychological or subjective *construct* through a set of *items* (questions or statements).

» **Construct:** the concept we want to measure
E.g., trust, satisfaction, motivation

» **Item:** a single question capturing part of that construct
E.g., "I trust the system to act in my best interest."

Wait, is the word "scale" used to mean two things?

Yes!

» **As a response instrument:**
The *format* for expressing responses (e.g., 1–7 agreement scale).

» **As a measurement instrument:**
The *set of items* combined to measure a construct.

Example: The term "scale" can refer to both:

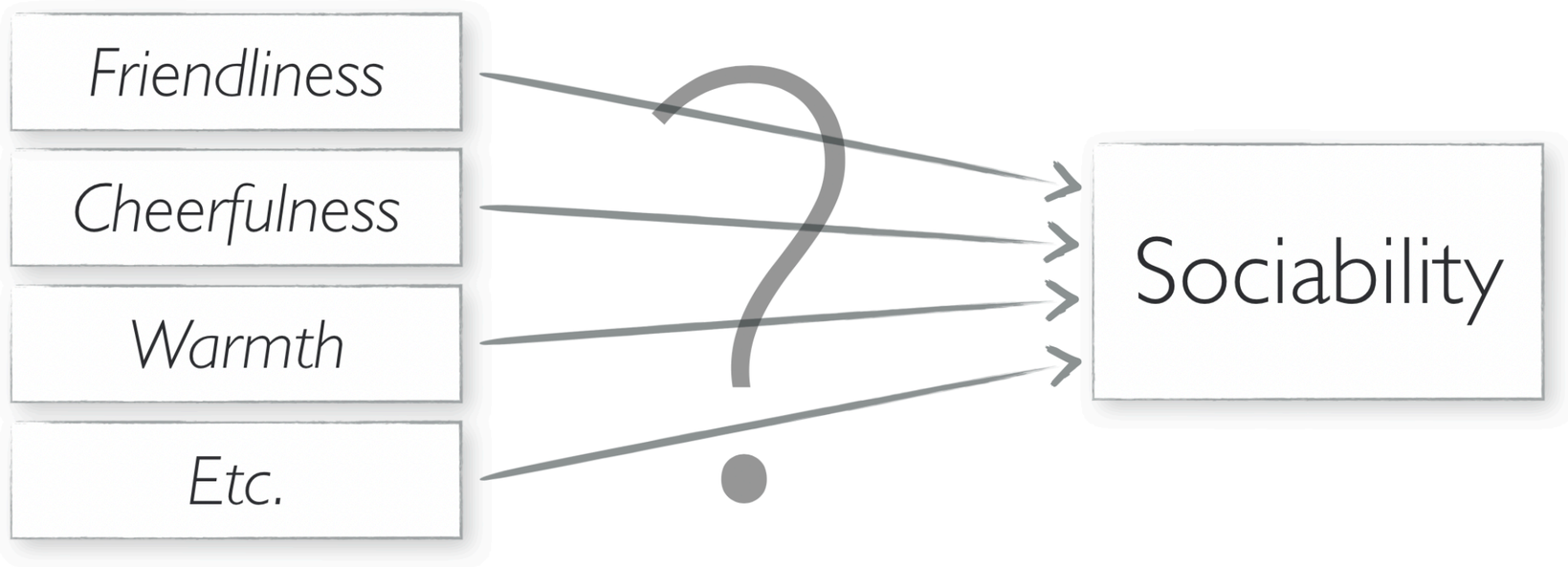
» The response scale, e.g., 7-point rating "scale" agreement format

» The measurement scale, e.g., a 10-item "scale" instrument assessing trust

Why are scales important?

An example: how can we measure *sociability*?

Too vague and multifaceted to be measured directly. Might be made up of sub-constructs, e.g., *friendliness*, *cheerfulness*, *warmth*, etc.



How do we create (or choose) scales?

1. **Define your construct**
   Identify what you want to measure (e.g., *trust*, *sociability*, *usability*).

2. **Generate items**
   Brainstorm possible components (e.g., *trust → credibility; usability → ease of use*).
   Use tools like WordNet to expand your item pool.

3. **Refine and test**
   Examine item relationships conceptually or statistically (e.g., factor analysis).

4. **Alternatively, adopt an existing validated scale**
   E.g., Usability Scales, Personality Item Pool (IPIP)

What does a validated scale look like? SUS, UEQ

	Strongly disagree				Strongly agree	
1. I think that I would like to use this system frequently					√	4
	1	2	3	4	5	
2. I found the system unnecessarily complex				√		1
	1	2	3	4	5	
3. I thought the system was easy to use		√				1
	1	2	3	4	5	
4. I think that I would need the support of a technical person to be able to use this system	√					4
	1	2	3	4	5	
5. I found the various functions in this system werw well integrated		√				1
	1	2	3	4	5	
6. I thought there was too much inconsistency in this system			√			2
	1	2	3	4	5	
7. I would imagine that most people would learn to use this system very quickly		√				1
	1	2	3	4	5	
8. I found the system very cumbersome to use				√		1
	1	2	3	4	5	
9. I felt very confident using the system					√	4
	1	2	3	4	5	
10. I needed to learn a lot of things before I could get going with this system		√				3
	1	2	3	4	5	

Total= 22 SUS Score= 22 * 2.5 = 55

	1	2	3	4	5	6	7		
annoying	O	O	O	O	O	O	O	enjoyable	1
not understandable	O	O	O	O	O	O	O	understandable	2
creative	O	O	O	O	O	O	O	dull	3
easy to learn	O	O	O	O	O	O	O	difficult to learn	4
valuable	O	O	O	O	O	O	O	inferior	5
boring	O	O	O	O	O	O	O	exciting	6
not interesting	O	O	O	O	O	O	O	interesting	7
unpredictable	O	O	O	O	O	O	O	predictable	8
fast	O	O	O	O	O	O	O	slow	9
inventive	O	O	O	O	O	O	O	conventional	10
obstructive	O	O	O	O	O	O	O	supportive	11
good	O	O	O	O	O	O	O	bad	12
complicated	O	O	O	O	O	O	O	easy	13
unlikable	O	O	O	O	O	O	O	pleasing	14
usual	O	O	O	O	O	O	O	leading edge	15
unpleasant	O	O	O	O	O	O	O	pleasant	16
secure	O	O	O	O	O	O	O	not secure	17
motivating	O	O	O	O	O	O	O	demotivating	18
meet expectations	O	O	O	O	O	O	O	does not meet expectations	19
inefficient	O	O	O	O	O	O	O	efficient	20
clear	O	O	O	O	O	O	O	confusing	21
impractical	O	O	O	O	O	O	O	practical	22
organized	O	O	O	O	O	O	O	cluttered	23
attractive	O	O	O	O	O	O	O	unattractive	24
friendly	O	O	O	O	O	O	O	unfriendly	25
conservative	O	O	O	O	O	O	O	innovative	26

What do I do when validated scales don't exist?

We develop our own by asking **good questions.**

Good questions are:

» Clear and unbiased

» Focused on firsthand experiences

» Interpreted consistently across respondents

» Easy to answer and record

Poorly designed questions lead to unreliable or invalid data.

How do I design "good" questions?[12]

1. Avoid bias and confusion

» No *leading* or *loaded* questions

» Avoid *double negatives* or *complex wording*

2. Ask about what people know

» Focus on *firsthand experiences*

» Avoid *hypotheticals* and *causal* explanations

3. Keep questions simple and focused

» Ask *one thing at a time*

» Avoid *unwarranted assumptions* or *hidden contingencies*

Example: "Would you like to be rich and famous?"

Better: *"**Q1:** Would you like to be rich?" "**Q2:** Would you like to be famous?"*

[12] See **Appendix** for full guidelines on designing "good" questions.

4. Ensure shared understanding

» Define key terms

» Present definitions *before* the question

» Specify time periods clearly

5. Structure questions for clarity

» Split complex questions into multiple items

» Use consistent wording and response order

» Clearly communicate expected responses

6. Make responding easy and consistent

» Keep layout simple and readable

» Train respondents or standardize administration

Example: "Would you say that you are very likely, fairly likely, or not likely to move out of this house in the next year?""

Better: *In the coming year, how likely are you to move to another home? Would you say very likely, fairly likely, or not very likely?*

Factor Analysis

So, I have good questions. How do I turn them into a ***scale*** *or* ***metric****?*

What Is factor analysis?

Goal: Reveal hidden dimensions (factors) that explain patterns in responses.

We use it to:

» **Discover structure** among related variables (data reduction)

» **Build scales** by grouping items that measure the same construct

Example: What underlying traits explain responses to "sociability" questions?

Factor analysis tells us which items "move together" and likely measure the same thing.

What are different types of factor analysis?

Type	Purpose	Question It Answers
Exploratory (EFA)	Find hidden structures among items	“What dimensions exist in my data?” "What are components of sociability?"
Confirmatory (CFA)	Test if hypothesized structure fits data	“Do these items really measure those factors?” "Do these items measure sociability well?"
Exploratory–Confirmatory	Combine discovery and testing	“Can I build and verify a usable scale?”

How do I conduct factor analysis?

1. **Collect data** by choosing relevant items
2. **Extract factors** (via PCA or similar)
3. **Decide how many to keep** (e.g., scree plot, eigenvalues > 1)
4. **Rotate** (e.g., varimax) to get simple, interpretable structure
5. **Interpret & refine** by identifying item–factor relationships
6. **Build scales** combining items by factor and test reliability

Exploratory analysis is iterative, involving test, refine, repeat until structure is clear.

How would I use exploratory vs. confirmatory analysis?
Principle: EFA finds the story; CFA tests if the story holds up.

Exploratory (EFA):

» Reveals *what factors exist* and *which items load* on them.

» Uses rotation (e.g., varimax) for clearer interpretation.

Confirmatory (CFA):

» Tests whether the proposed structure *fits* the data.

» Constrains weak loadings to zero, compares models statistically.

Exploratory–Confirmatory:

» Combine both: discover structure, then confirm with new data.

What is an example factor analysis problem?

Imagine that we are interested in measuring factors that might affect people's decisions about buying a car.

We design a questionnaire with a number of items that we think will be relevant: price, safety, exterior appearance, space/comfort, technology, after sales service, resale value, fuel type, fuel efficiency, color, maintenance, test drive, product reviews, testimonials.

How important is the following factors in your decision to purchase?

Price	Not important	1	2	3	4	5	Important
Safety	Not important	1	2	3	4	5	Important
Exterior appearance	Not important	1	2	3	4	5	Important
Space/comfort	Not important	1	2	3	4	5	Important
Technology	Not important	1	2	3	4	5	Important
After sales service	Not important	1	2	3	4	5	Important
Resale value	Not important	1	2	3	4	5	Important
Fuel type	Not important	1	2	3	4	5	Important
Fuel efficiency	Not important	1	2	3	4	5	Important
Color	Not important	1	2	3	4	5	Important
Maintenance	Not important	1	2	3	4	5	Important
Test drive	Not important	1	2	3	4	5	Important
Product reviews	Not important	1	2	3	4	5	Important
Testimonials	Not important	1	2	3	4	5	Important

Given m factors and n observed variables:

$$X_1 = \lambda_{11} F_1 + \lambda_{12} F_2 + \ldots + \lambda_{1m} F_m + e_1$$

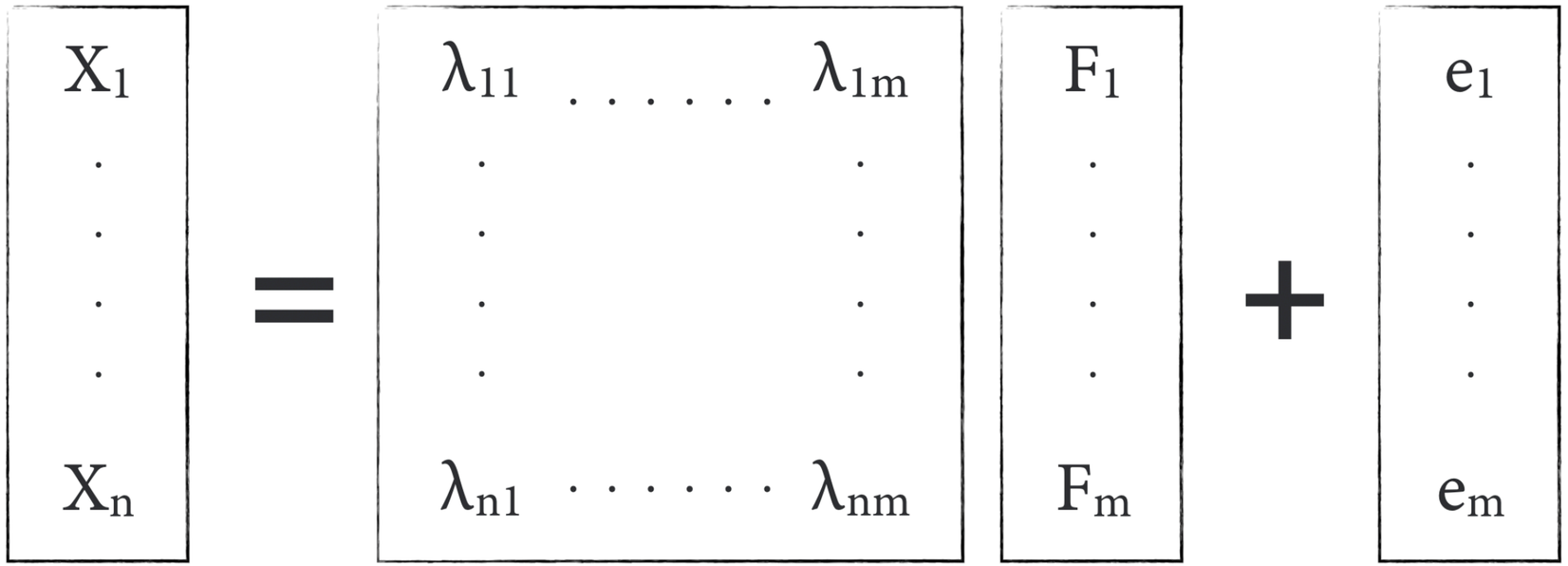
$$X_2 = \lambda_{21} F_1 + \lambda 22 F_2 + \ldots + \lambda 2m F_m + e_3$$

$$\ldots$$

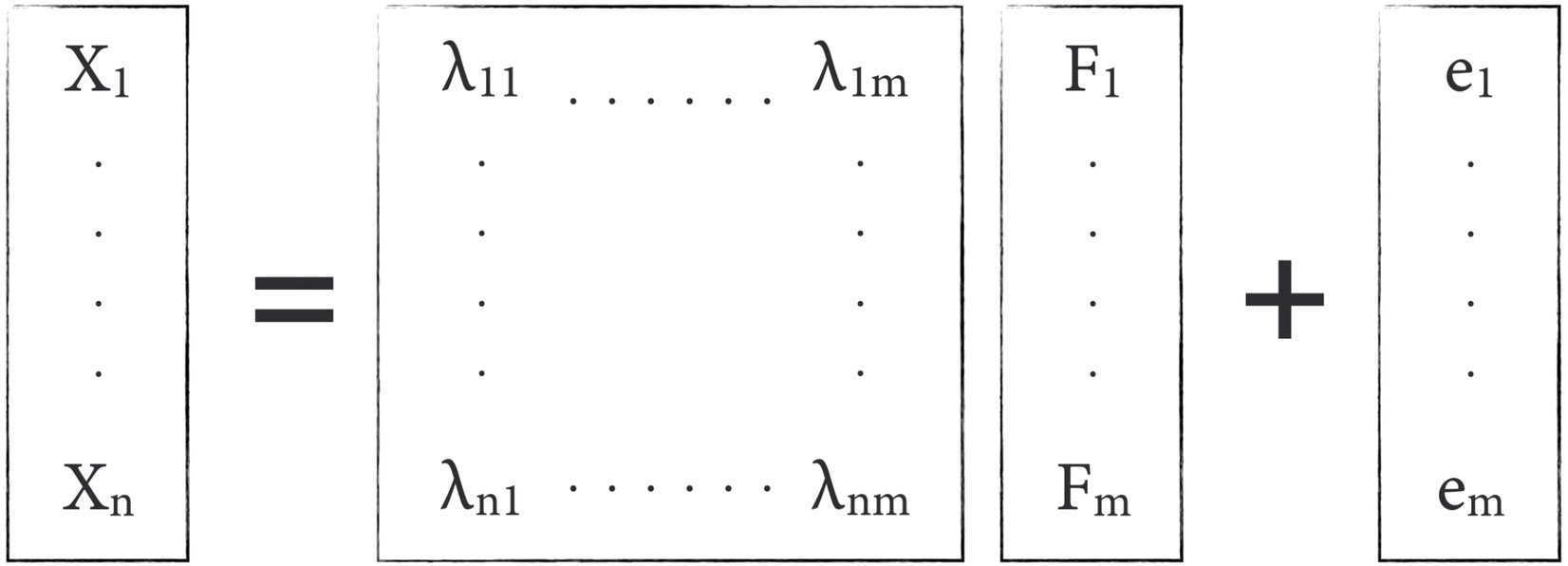
$$Xn = \lambda n1 F1 + \lambda n2 F2 + \ldots + \lambda nm Fm + en$$

In matrix notation:

$$X_{n\times 1} = \Lambda_{n\times m} F_{m\ times 1} + e_{n\times 1}$$



How do we interpret the factor matrix?



What is factor rotation?

Definition: Factor rotation is a statistical technique that allows us to make more clear-cut decisions by spreading variability more evenly among factors by redefining factors to force loadings to be very high (-1 or 1) or very low (0).

There are different methods of factor rotation. We will use *varimax*, which maximizes squared loading variance across variables (sum over factors).

Let's try it out![8] [9]

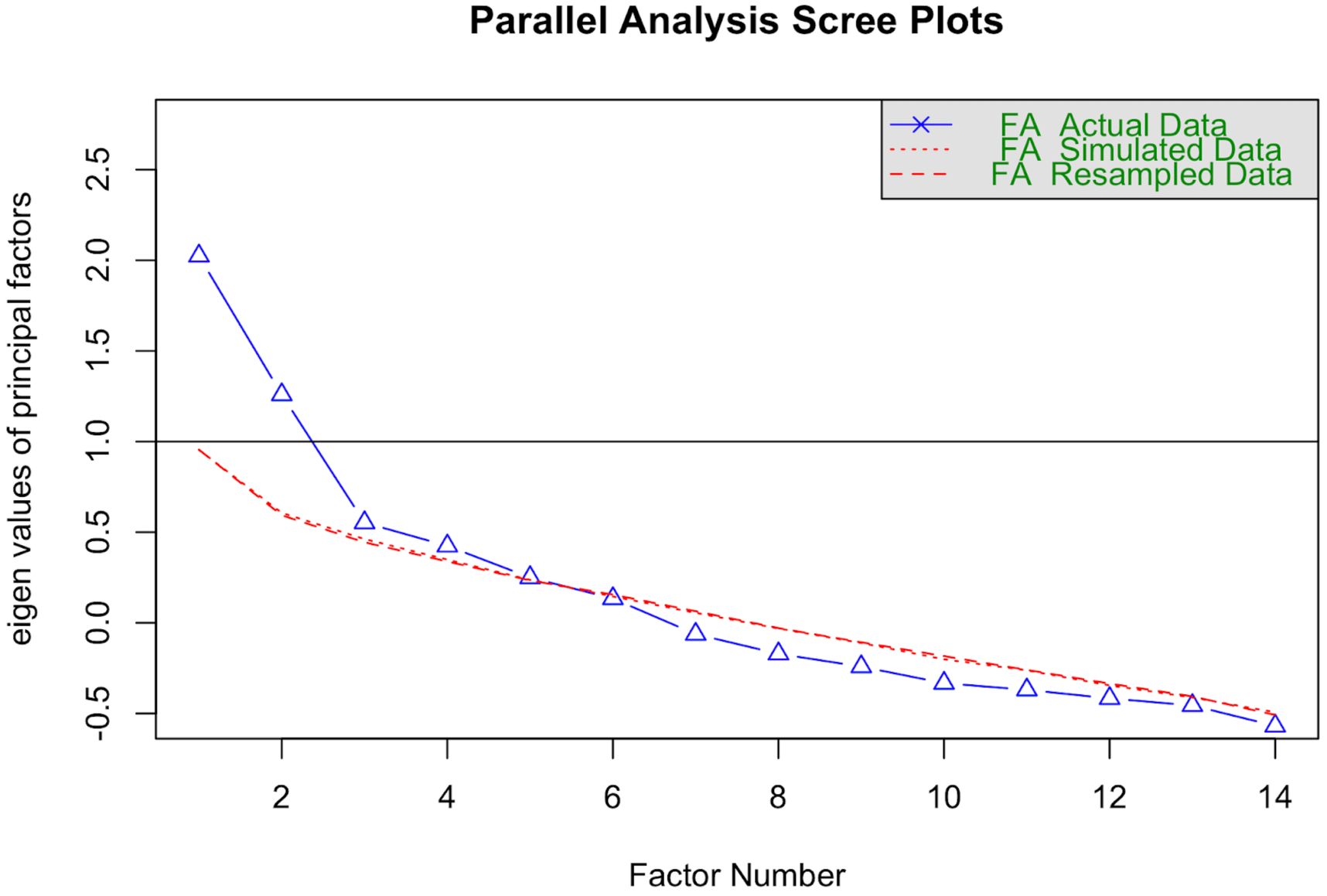
[8] We'll use R.

[9] We will use same data from PromptCloud.

Step 1. Determine the number of factors using PCA

```
install.packages("psych")
library(psych)
library(readr)
EFA <- read_csv("EFA.csv")
pa = fa.parallel(EFA, fm = 'minres', fa = 'fa')
```

This will produce what is called a Scree plot that will plot eigenvalues on the Y axis and number of factors on the X axis.


Parallel Analysis Scree Plots
FA Actual Data
FA Simulated Data
FA Resampled Data
eigen values of principal factors
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
2
4
6
8
10
12
14
Factor Number

How do we determine the number of factors?

There are two methods:

1. **Kaiser Criterion**:[10] take eigenvalues that are larger than 1.

2. **Scree test**:[11] find point of inflection and consider the factors up to the leveling off.

---

[10] Kaiser (1960). The application of electronic computers to factor analysis.

[11] Cattell (1966). The Scree test for the number of factors.

Step 2: Factor rotation

Calculate loadings for each variable on each factor:

$$corr(F_i, X_j) = \lambda_{ji}$$

Apply factor rotation to spread the variability evenly among variables:

```
fit = fa(data,nfactors = 3,rotate = "varimax",fm="minres")
```

Visualize the factor matrix:

```
print(fit$loadings,cutoff = 0.3)
```

This will print out the following factor matrix:

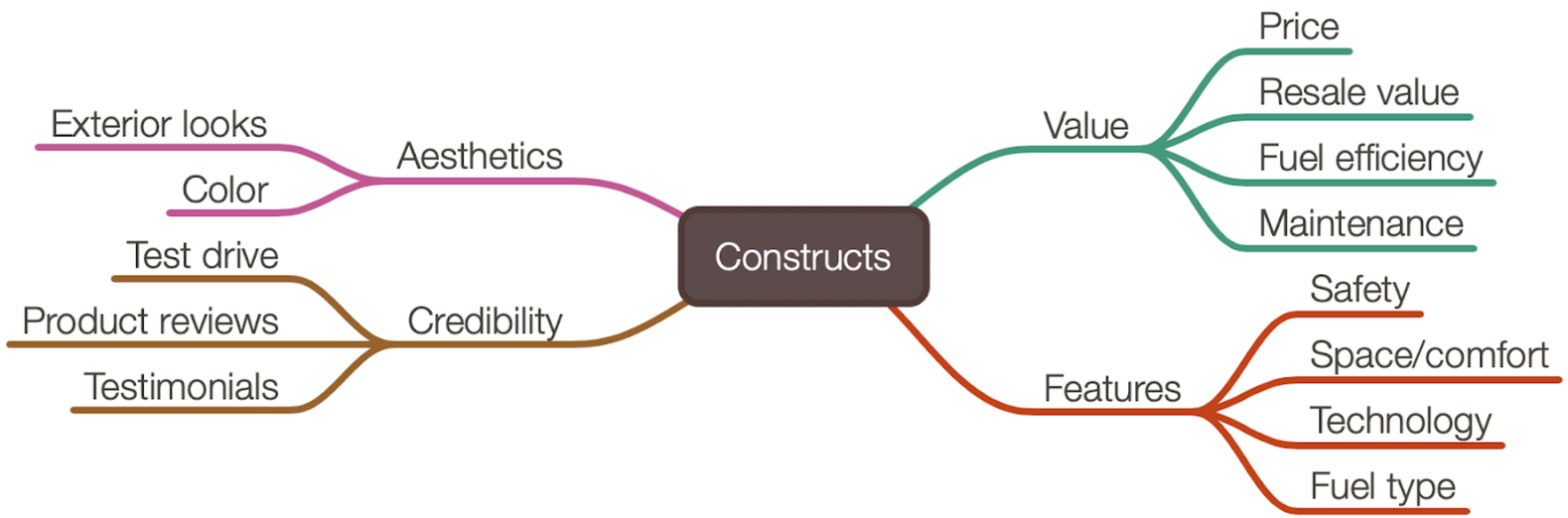
```
Loadings:
                    MR1    MR2    MR3
Price                0.444
Safety                      0.311
Exterior_Looks
Space_comfort               0.832
Technology                  0.342
After_Sales_Service         0.460
Resale_Value         0.599
Fuel_Type                   0.573
Fuel_Efficiency      0.655
Color                0.464
Maintenance          0.668
Test_drive                         0.328
Product_reviews      0.424
Testimonials                       0.742
```

We can iteratively interpret and recalculate:

```
Loadings:
                    MR1    MR2    MR3    MR4
Price               0.535
Safety                     0.356
Exterior_Looks                           -0.544
Space_comfort              0.753
Technology                 0.349
After_Sales_Service        0.528
Resale_Value        0.724
Fuel_Type                  0.557
Fuel_Efficiency     0.492
Color                                     0.706
Maintenance         0.603
Test_drive                        0.407
Product_reviews     0.336         0.429
Testimonials                      0.677
```

Step 3: Scale construction

We inspect all factors and items that load to them:



To create a scale, we combine the items that load to that scale:

```
scale_value = cbind(EFA$Price,EFA$Resale_Value,EFA$Fuel_Efficiency,EFA$Maintenance)
```

Step 4: Test scale reliability

Recap: Most commonly used measure of scale reliability is Cronbach's α.

Cronbach's alpha — Internal consistency

» $\alpha \geq .9$ — Excellent

» $.9 > \alpha \geq .8$ — Good

» $.8 > \alpha \geq .7$ — Acceptable

» $.7 > \alpha \geq .6$ — Questionable

» $.6 > \alpha \geq .5$ — Poor

» $.5 > \alpha$ — Unacceptable

To calculate Cronbach's α:

```
alpha(scale_value, na.rm = TRUE)
```

This will produce:

```
Reliability analysis
Call: alpha(x = scale_value, na.rm = TRUE)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean   sd median_r
      0.65      0.67    0.61      0.34   2 0.055  3.9 0.61     0.31

 lower alpha upper     95% confidence boundaries
0.54 0.65 0.76

 Reliability if an item is dropped:
                raw_alpha std.alpha G6(smc) average_r S/N alpha se   var.r
scale_value          0.60      0.62    0.53      0.35 1.6    0.067 0.00556
Resale_Value         0.54      0.55    0.45      0.29 1.2    0.081 0.00015
Fuel_Efficiency      0.60      0.64    0.55      0.37 1.8    0.061 0.00467
Maintenance          0.55      0.59    0.50      0.33 1.5    0.072 0.00233
                med.r
scale_value      0.33
Resale_Value     0.30
Fuel_Efficiency  0.38
Maintenance      0.33
```

To use the scale, we will average items out:

```
average <- rowMeans(scale_value)
scale_numerical <- data.frame(Average = average)
print(scale_numerical)
```

```
# Load libraries & data
install.packages("psych")
library(psych)
library(readr)
EFA <- read_csv("EFA.csv")

# Calculate factor matrix using Parallel Analysis
pa = fa.parallel(EFA, fm = 'minres', fa = 'fa')

# Factor rotation
fit = fa(data,nfactors = 3,rotate = "varimax",fm="minres")
print(fit$loadings,cutoff = 0.3)

# Calculate scale reliability
scale_value = cbind(EFA$Price,EFA$Resale_Value,EFA$Fuel_Efficiency,EFA$Maintenance)
alpha(scale_value, na.rm = TRUE)

# Use the scale
average <- rowMeans(scale_value)
scale_numerical <- data.frame(Average = average)
print(scale_numerical)
```

Appendix

Principles of Good Question Design

When validated scales don't exist, we must develop our own by writing **good questions** that are clear, unbiased, and consistent.

The following principles provide guidance and examples...

Principle 1: *Avoid leading or loaded questions — Don't bias the response with emotionally or morally charged wording.*

Example

Don't you agree that social workers should earn more money than they currently earn?

Better

Do you think social workers are fairly compensated for their work?

Why? Leading wording pressures respondents toward one answer.

Principle 2: *Avoid double negatives — Negations within negations are confusing and reduce accuracy.*

Example

Do you agree or disagree that teachers should not be required to supervise students during recess?

Better

Should teachers be required to supervise students during recess?

Why? Positive phrasing improves clarity and comprehension.

Principle 3: *Ask about firsthand experience — People can describe what happens to them, not what happens generally.*

Example

How much crime occurs in your neighborhood?

Better

Have you personally experienced or witnessed crime in your neighborhood in the past year?

Why? Respondents are better reporters of their own experiences than of community trends.

Principle 4: *Avoid hypothetical questions — People are unreliable at predicting future or imagined behavior.*

Example

If you won the lottery, how would you spend the money?

Better

How did you spend the last large sum of unexpected money you received?

Why? Questions about real experiences yield more valid answers.

Principle 5: *Avoid asking about causality — People rarely know why events occur or why they act as they do.*

Examples — Poor

Were you limited in daily activities because of your back problem?
Why didn't you vote in the last election?

Better

What factors made it difficult for you to vote in the last election?

Why? Causality is complex; ask about conditions or contributing factors instead.

Principle 6: *Avoid asking for solutions to complex problems — People lack enough expertise to propose solutions in surveys.*

Example

What should the government do to end homelessness?

Better

Which approaches to addressing homelessness do you support or oppose?

Why? Surveys should capture attitudes or evaluations, not policy design.

Principle 7: *Ask one question at a time — Avoid combining multiple concepts in a single item.*

Example

Would you like to be rich and famous?

Better

Would you like to be rich?
Would you like to be famous?

Why? Two distinct ideas can produce conflicting responses.

Principle 8: *Avoid unwarranted assumptions — Don't include built-in assumptions that may not hold true.*

Example

Should the organization reduce paperwork by hiring more administrators?

Better

Should the organization reduce paperwork?
Should it hire more administrators?

Why? Breaking assumptions apart ensures clarity and fairness.

Principle 9: *Avoid hidden contingencies — Don't assume respondents share specific contexts or habits.*

Example

How often do you interact with people on Facebook?

Better

Which online or in-person platforms do you use to interact with others, and how often?

Why? Not everyone uses Facebook; questions must apply to most respondents.

Principle 10: *Use clear and shared language — Choose words that all respondents understand in the same way.*

Example

How do you feel about the new policy?

Better

How satisfied are you with the new office attendance policy?

Why? Precise, concrete wording avoids ambiguity.

Principle 11: *Define specialized terms — Provide clear definitions for any term that could be interpreted differently.*

Example

How many times have you seen or talked with a medical doctor about your health?

Better

Include visits to psychiatrists, ophthalmologists, or any professional with a medical degree.
In the past 12 months, how many times have you seen or talked with such a doctor?

Why? Definitions standardize interpretation across respondents.

Principle 12: *Provide definitions before the question — Clarify terms before asking so respondents use them while answering.*

Example

How many days in the past week have you done any exercise?
When answering, include walking or housework.

Better

The next question is about exercise.
Include walking, household chores, or job-related activity.
Using this definition, on how many days in the past week did you exercise?

Why? Providing definitions first reduces confusion and re-interpretation.

Principle 13: *Specify time frames clearly — Make the reference period unambiguous.*

Example

Do you exercise regularly?

Better

In the past 7 days, how many times have you exercised for at least 30 minutes?

Why? Precise time references improve recall and comparability.

Principle 14: *Split complex topics into multiple questions — Don't overload a single question with multiple dimensions.*

Example

How satisfied are you with your pay and benefits?

Better

How satisfied are you with your pay?
How satisfied are you with your benefits?

Why? Separate items reveal distinct patterns and reduce confusion.

Principle 15: *Use multiple items to measure a concept — Capture constructs using several related questions.*

Example

To measure trust, ask about honesty, reliability, and competence.

Why? Multiple items improve reliability and allow averaging across responses.

Principle 16: *End with the question — Place response options at the end, not in the middle.*

Example

Would you say that you are very likely, fairly likely, or not likely to move in the next year?

Better

In the coming year, how likely are you to move?
(Very likely / Fairly likely / Not very likely)

Why? Clear separation improves readability and comprehension.

Principle 17: *Clarify expected answers — Make response format and expectations clear.*

Example

When did you move to this community?

Possible answers:

When I was sixteen. / Right after I was married. / In 1953.

Better

In what year did you move to this community?

Why? Specify the type of response expected (year, number, category).

Principle 18: *Specify number of responses allowed — Tell respondents how many options they can select.*

Example

What made you choose this brand?
(Select all that apply.)

Why? Instructions prevent incomplete or over-selected answers.

Principle 19: *Simplify layout and instructions — Reduce cognitive and visual effort to minimize errors.*

Tips

» Use clear fonts and spacing

» Provide consistent question numbering

» Keep response options visually aligned

Why? Accessibility and readability support accurate data entry and interpretation.

Principle 20: *Orient respondents to the task — Help people answer in a consistent, focused way.*

How

» Provide clear instructions at the start

» Offer a short practice question if appropriate

» Train interviewers or respondents when needed

Why? Consistency across respondents improves data quality.

Summary: *Designing* ***good questions*** *ensures that new scales are valid and reliable.*

» Avoid bias and confusion

» Focus on firsthand experience

» Use clear, shared language

» Keep structure simple and consistent

» Support respondents in answering accurately

Quick Reference: Principles of Good Question Design

#	**Principle**	**Core Idea**
1	Avoid leading questions	Remove biasing language
2	Avoid double negatives	Keep wording positive
3	Ask about firsthand experience	Focus on what people know
4	Avoid hypotheticals	Ask about real experiences
5	Avoid causality	Don’t ask “why” people act
6	Avoid complex solutions	Focus on opinions, not policies
7	Ask one question at a time	No double-barrels
8	Avoid unwarranted assumptions	Don’t imply relationships
9	Avoid hidden contingencies	Ensure relevance to all respondents
10	Use shared language	Choose clear, concrete words

#	Principle	Core Idea
11	Define specialized terms	Clarify meaning of key terms
12	Provide definitions first	Present definitions before questions
13	Specify time frames	Anchor recall periods
14	Split complex topics	Ask separate questions
15	Use multiple items	Measure constructs with several items
16	End with the question	Keep response options at the end
17	Clarify expected answers	Tell respondents the response format
18	Specify number of responses	Indicate if multiple answers allowed
19	Simplify layout	Make survey easy to read and navigate
20	Orient respondents	Train or guide for consistency