



VOLUME 1

# Correlations

A resource for action and consensus in clinical measurement

# Our contributors

The 32 people below shaped Volume 1 through review, discussion and data. Listed alphabetically by first name.

**Aaron Yaras**

Ionis Pharmaceuticals

**Abi Williams**

Adelphi Values

**Alexandra Lauer**

Boehringer Ingelheim

**Aline Frick**

TiliaHealth

**Amy M Jones**

Novo Nordisk

**Brandon Foster**

Personal capacity

**Bryce B. Reeve**

Population Health Sciences, Duke University

**Catarina Lopes Dias**

University of Graz

**Dara O'Neill**

Patient Centered Solutions, IQVIA

**David A. Andrae**

BioMarin

**Don Bushnell**

Thermo Fisher Scientific

**Dorothee Oberdhan**

University of Sunderland

**Elizabeth (Lizzie) Collins**

Clarivate

**Emmanuel Streel**

Altoida, Inc.

**Franco Di Cesare**

Leoben Research Aurora

**Jacque Mogle**

RTI International

**Jenya Antonova**

Compass Strategy and Research, Inc.

**Jessie P Bakker**

Candlagan Consulting LLC

**Jonathan Comins**

Sports Traumatology, Copenhagen University Hospital

**Julio César Sánchez-Arévalo**

Universidad FUCS

**Katerina Stravoravdi**

Clinical Research Unit, Papageorgiou General Hospital, School of Medicine, AUTH

**Kristine Ogden**

Otsuka Pharmaceutical Development & Commercialization, Inc.

**Kshawna Askew**

Independent

**Marta Surlej**

PHUSE RBQM Working Group

**Meg Fluharty**

Acaster Lloyd Consulting

**Nace Mikus**

TiliaHealth

**Oren Meyers**

GSK

**Prathiba Batley**

GSK

**Sarah Clifford**

Sprout Health Solutions

**Sarah M. Donelson**

Bristol Myers Squibb

**Theresa Coles**

Duke University

**Tim Pickles**

Leeds Beckett University

WITH THANKS TO OUR COLLABORATORS

# Corporate Collaborators

The organizations below back Volume 1 with people, review and data.

|                                                                                                                        |                                                                                                                   |                                                                                                                      |                                                                                                         |                                                                                                                                              |                                                                                                                      |
|------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|
|  <p>Acaster Lloyd Consulting</p>      |  <p>Adelphi Values</p>           |  <p>Patient-Centered Outcomes</p>  |  <p>Altoida</p>      | <p>In collaboration with</p>  <p>Boehringer Ingelheim</p> |  <p>Candlagan Consulting LLC</p>  |
|  <p>Compass Strategy and Research</p> |  <p>Duke University</p>          |  <p>ECNP Clinical Outcomes TWG</p> |  <p>GSK</p>          |  <p>Ionis Pharmaceuticals</p>                             |  <p>Leeds Beckett University</p>  |
|  <p>Leoben Research Aurora</p>       |  <p>Sprout Health Solutions</p> |  <p>Thermo Fisher Scientific</p>  |  <p>TiliaHealth</p> |  <p>Universidad FUCS</p>                                 |  <p>University of Sunderland</p> |

Is your organization missing? If your team contributed data, review or expertise to Volume 1 and is not listed yet, let us know and we'll add your logo to the next printing.



## Why this resource exists, and who it is for



### WHY

There are many psychometric methods, implementations of the methods, and interpretations of the results. This wealth of options leads to confusion and inconsistency. This series aims to bring cohesion and consensus to our field



### WHO

This series is for those who are learning - students and new investigators - who may need pointing in the right direction or grounding. It is also for experienced scientists who need a reference they can point to and to increase consistency between them. It is written by people who do this work daily, for people who will have to defend it.



### WHERE

Broadly, the series should be used in the development of regulatory documentation and health psychometric more broadly. For this volume specifically, it should be used at the points where a correlation is needed for decision making: the analysis plan, the validation report, the submission dossier, and the review of someone else's analysis.



### WHAT

Volume 1 covers correlations. Later volumes take the same shape for the other methods or properties.



# Glossary

**Between-subject correlation** — Association between individuals' average levels of two variables. Reflects stable differences between people.

**Within-subject correlation** — Association between deviations from each person's own mean across occasions. Reflects co-variation for an individual over time.

**Total correlation** — A standard correlation across all observations, ignoring nesting. Conflates both between subjects and within subject correlations; not interpretable alone.

**Nesting** — A structure in which observations belong to a higher unit (timepoints within a person). Violates the independence assumption.

**Person-mean centering** — Subtracting each person's mean from their observations to separate within- from between-subject variability.

**Intraclass correlation (ICC)** — Proportion of total variance from stable between-subject differences; here, the degree of clustering within people.

**Ecological fallacy** — Inferring individual-level dynamics from group-level associations (Robinson, 1950; Curran & Bauer, 2011).

**Simpson's Paradox** — Statistical phenomenon in which a trend observed within a patient, reverses or disappears when between-patients trend is analyzed.



# Glossary

**Dependence** — Two observations are dependent when knowing one changes what you'd expect for the other. Two visits from the same patient, unlike two different patients.

**Continuous variable** — A variable that can take any infinite number of real values within a range.

**Ordinal variable** — A variable with categories that have a logical order but the exact distance between the categories is unknown.

**Binary variable** — A variable with two categories only.

**Corrected (item-total correlation)** — « Corrected » means the item is excluded from the total score it's compared against, so it isn't partly correlated with itself.

**Clinician Global Impression of Severity (CGIS)** — A single item questionnaire asking the clinician to rate the severity of a patient's condition, usually at the time of assessment.

**Clinician Global Impression of Change (CGIC)** — A single item questionnaire asking the clinician to rate the improvement or worsening in a patient's condition since the start of a study.

**Patient Global Impression of Severity (PGIS)** — A single item questionnaire asking the patient about their condition, usually referenced over the last week.

**Patient Global Impression of Change (PGIC)** — A single item questionnaire asking the patient to rate the improvement or worsening in their condition since the start of a study.

**Lord's Paradox** — The longitudinal analogue of Simpson's Paradox: adjusting for baseline can flip the direction of a change effect.



## TOPIC 1

# What is a correlation coefficient?

The coefficient, its assumptions, and where it shows up in clinical outcome assessment



# What is a correlation coefficient?

A correlation coefficient is a single number that captures how strongly two variables move together, and in which direction.

$$r = \text{cov}(X, Y) / (s_X \cdot s_Y)$$

Covariance of X and Y divided by the product of the standard deviations of X and Y

- Sign gives the direction: positive means both variables tend to increase or decrease together; negative means one tends to increase as the other decreases. Sign doesn't give the strength, a correlation of -0.9 is as strong as a correlation of 0.9
- Magnitude gives the strength: 0 is no linear association, 1 is a perfect one. Within this range, the magnitudes are classified as negligible, small, moderate, and strong ([threshold benchmarks](#)).
- Standardized and bounded: correlation coefficient ranges from -1 and +1, so they are comparable across different situations (measures and item pairs across one measure).
- Written r, not "correlation": r is one specific coefficient (Pearson's, used for continuous data). Other types of correlation coefficient suit other types of data (see [Topic 3](#)).



*Illustrative scatter at four values of r.*



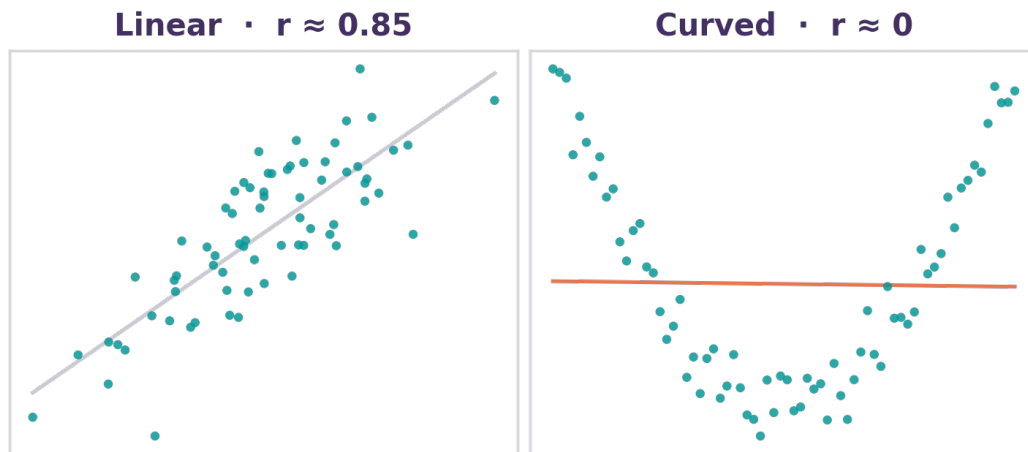
## KEY INSIGHT

- A correlation is a standardized effect size: its magnitude tells you how strong the association is.
- Judge correlation coefficient by size, not by whether it statistically differs from zero.



# Linearity and the assumptions behind $r$

Pearson's coefficient ( $r$ ) measures the strength of the linear relationship only, so always plot the data before trusting it.



*Same underlying relationship, very different  $r$ .*

Ordinal, skewed, or monotonic-but-non-linear data → use **Spearman's  $\rho$**  (see [Topic 3](#) for more details on Spearman and other correlation coefficients existing).

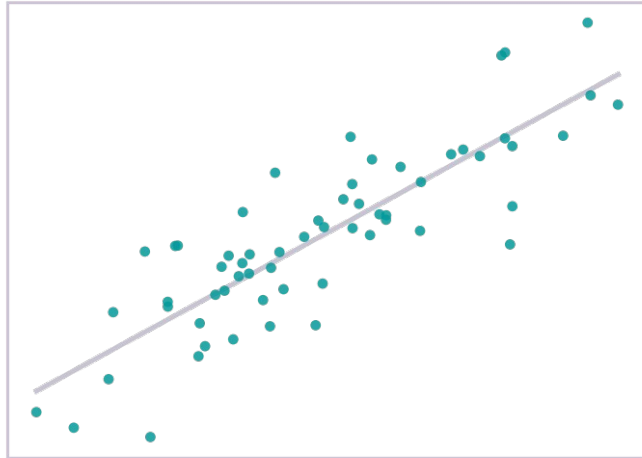
## ASSUMPTIONS FOR PEARSON $r$

- ✓ Linear relationship
- ✓ No dominant outliers
- ✓ Unrestricted (adequate) range
- ✓ Continuous, interval-level data
- ✓ Even spread of points (homoscedasticity)
- ✓ Roughly bivariate-normal shape (for p-values and confidence intervals(CIs))

*Next: what each assumption looks like.*

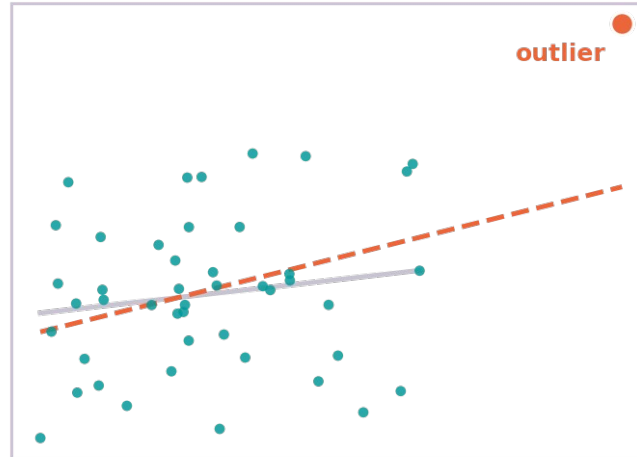


## What the assumptions look like (1 of 2)



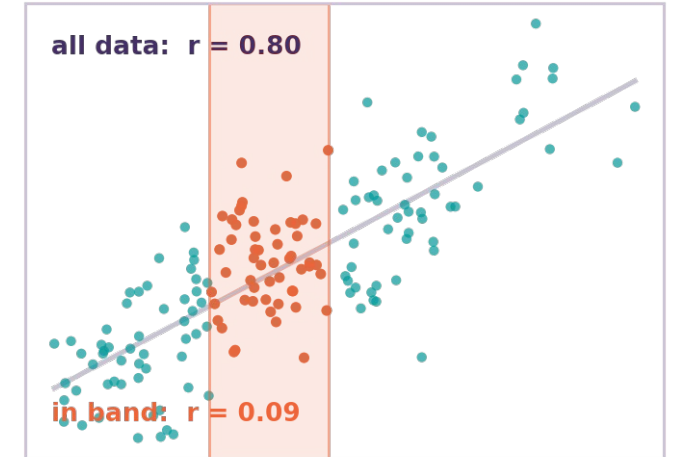
### Linear relationship

The scatterplot formed by the data follow a straight line. The Pearson coefficient ( $r$ ) can capture this relationship



### No dominant outliers

One extreme point can create (or hide) a correlation. Check its influence before trusting  $r$ .

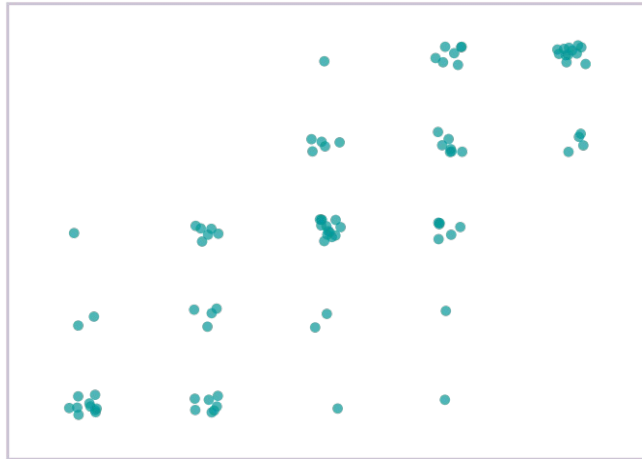


### Unrestricted range

Narrowing the range shrinks  $r$ . Here it falls from 0.80 across all data to near zero inside the band ([restricted range](#)).

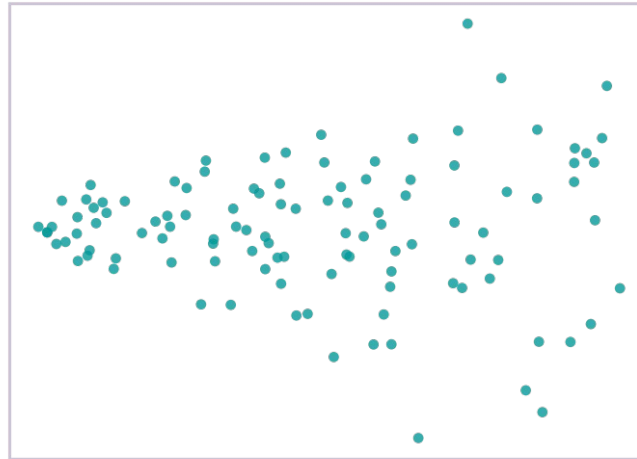


## What the assumptions look like (2 of 2)



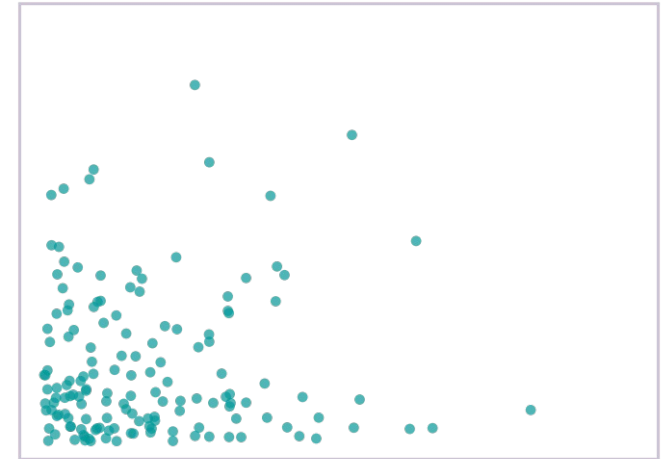
### Continuous data

Few discrete levels (e.g., a 5-point item) suit Spearman or polychoric correlations ([Topic 3](#)).



### Even spread

Spread should stay roughly constant. A fan shape signals heteroscedasticity and weakens inference.



### Roughly normal shape

Skew and heavy tails distort p-values and confidence intervals, though not the estimate itself.



# Correlation is not causation

A correlation shows that two measures vary together. It does not show that one causes the other.

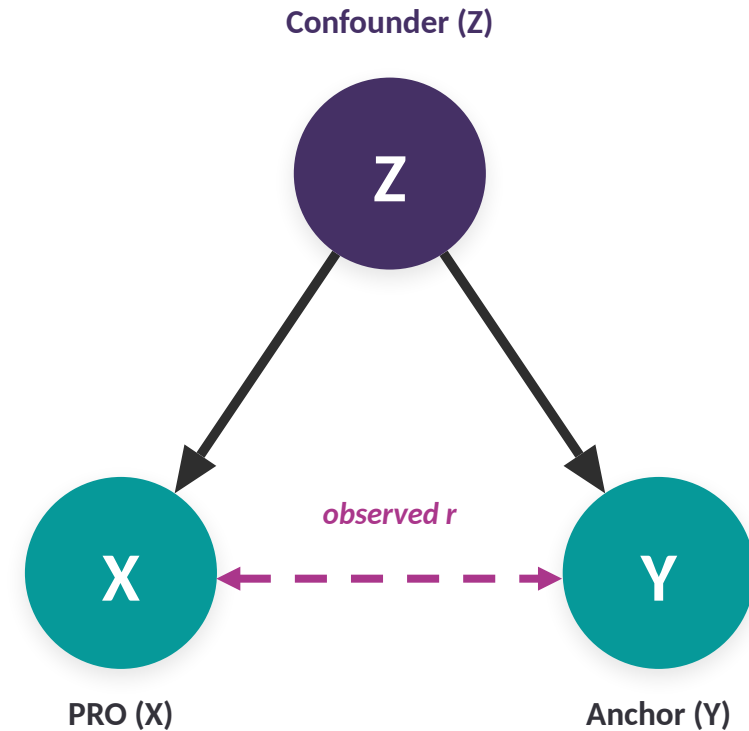
Correlations could be caused by:

- Confounding: a third variable drives both.
- Reverse causation: the effect may run the other way.
- Chance or selection: coincidence, or how the sample was drawn.



## IN PRACTICE

In evaluation of Clinical Outcome Assessments (COAs) data, a strong PRO–anchor (patient-reported outcome–anchor) correlation supports a shared construct (validity evidence). It does not prove that changing one changes the other.





## Correlation is not causation (2 of 2)

### Confounding

**Situation:** A strong correlation between poor scores on a psychometric memory test (such as the Mini-Mental State Evaluation, MMSE) and high scores on a gait disorder rating scale in elderly patients is observed.

**Wrong Conclusion:** « *Memory loss causes disorientation, which leads to a loss of balance.* »

**Correct conclusion:** Confounding, underlying pathology causes both.

### Reverse causation

**Situation:** A strong correlation between a high score on a health anxiety inventory and an irritable bowel syndrome severity is observed.

**Wrong Conclusion:** « *These patients' anxiety is psychosomatic and causes their intestinal pain.* »

**Correct conclusion:** The causality is reversed.

### Chance or selection

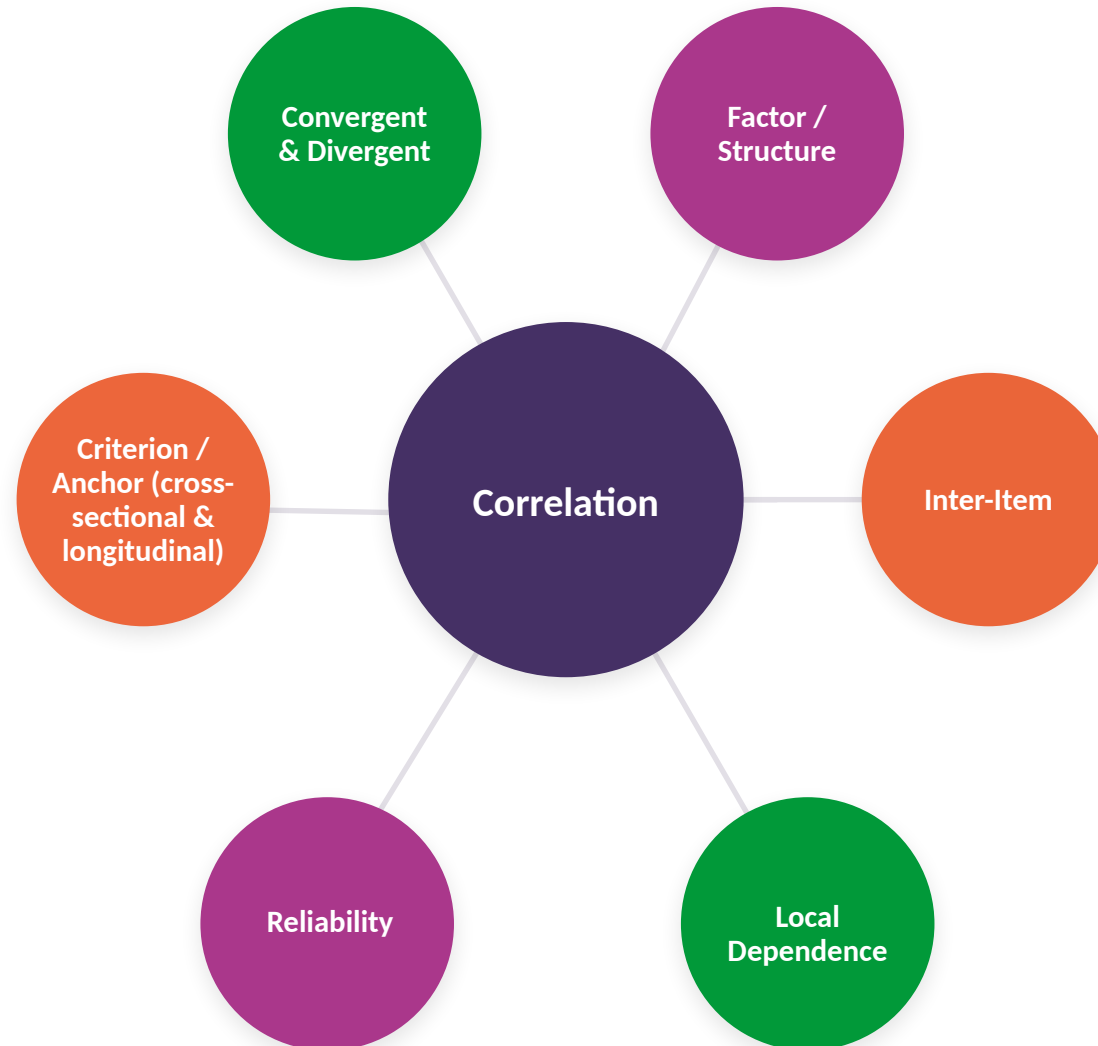
**Situation:** A strong correlation is found between a severity score of schizophrenia and the likelihood of developing cardiovascular disease. The study is conducted in a hospital emergency department.

**Wrong Conclusion:** « Schizophrenia causes cardiovascular disease. »

**Correct conclusion:** Looking only at hospitalized patients, an artificial correlation is created.



## Correlation underpins many elements of COA





# How correlation supports evaluation of psychometric properties of a COA instrument

## (1 of 2)



### Inter-item correlations

Correlations among the items in a scale show whether they belong together. Moderate, positive values suggest items tap into the same construct; a very high correlation coefficient may signal redundancy, and a near-zero correlation coefficient may indicate that the items do not contribute to the same scale (i.e., construct). Inter-item relies on [reliability](#).



### Factor analyses and item structure

The correlation matrix is the raw material for factor analysis (can also be a covariance matrix). Its pattern reveals the latent structure: how many factors are present in the data, and which items are associated with each factor (EFA, exploratory factor analysis and CFA, confirmatory factor analysis). A complete understanding of the latent structure requires other analyses ([latent variables](#)).



### Convergent & divergent validity

Validity evidence rests on the size of correlations with other measures: strong links with the same construct (convergent) and weaker links with unrelated ones (divergent) (Campbell & Fiske, 1959). Other measures used must be validated themselves.



# How correlation supports psychometric properties of a COA instrument (2 of 2)



## Criterion correlation / Anchor

Correlating a COA with an external anchor or criterion (such as a global rating) supports interpretation and underpins responsiveness and meaningful-change work. The anchor must itself be valid, reliable and at least moderately related to the concept of interest.



## Test-retest reliability

The correlation or intraclass correlation ([ICC](#)) between assessment on two occasions in stable patients is an index of reproducibility. A high value means the score stays consistent when the underlying state has not changed.



## Local dependence

In factor analyses, once the latent trait has been identified, item residuals should be roughly uncorrelated. Leftover residual correlations signal local dependence, which can inflate reliability estimates and distort scoring (Chen & Thissen, 1997).



## Key takeaways

1

A correlation is a standardized effect size for association: direction from the sign, strength from the magnitude, range of  $-1$  to  $+1$  scale.

2

Pearson's  $r$  captures linear association only. Plot the data first and choose the right coefficient (e.g. Spearman for ordinal data).

3

Association is not causation: confounding, reverse causation, and chance can each produce a correlation.

4

Across COA work, correlations underpin reliability, structure, and validity evidence; the research question determines what you compute.



## References and further reading

### Slide-specific references

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *Am Sociol Rev*, 15, 351-357;

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annu Rev Psychol*, 62, 583-619.

### Volume-wide further reading:

**Slide 13** Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105.

**Slide 14** Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *J Educ Behav Stat*, 22(3), 265-289.

**Foundations** Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences* (3rd ed.). Erlbaum;

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill;

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*, 24(3), 69-71;

FDA (2009). PRO Measures: Use in Medical Product Development to Support Labeling Claims.



Correlations

TOPIC 2

# Methods for correlation-based questions

Choosing the right correlation for the question you are asking



# Three methods, three questions

The data structure and the research question determine the type of correlation to compute.



## Cross-sectional

*Analysis of correlation at one timepoint*

A single timepoint is available for all patients.

*Do patients who differ on X also differ on Y?*

**Application:** Construct validity • usually strong correlation  $r$  for related measures (X and Y)



## Change / longitudinal

*Analysis of change over time*

Multiple timepoints are available for all patients.

*Do patients who change on X also change on Y?*

**Application:** Responsiveness • usually weaker  $r$  because of compound error



## Within & between

Repeated measures, nested in patients.

*How much association is within vs between patients?*

Separates the two ([Within Between](#))

**Cross-sectional and longitudinal methods contribute to the validation argument:** cross-sectional correlations show the instrument captures the concept; change correlations support that it can detect change over time and within patients.



## Correlating two measures at one timepoint

A cross-sectional correlation asks whether two measures increase and decrease together, among patients at a single timepoint. It reflects differences between patients at a given timepoint, not changes within them over time.

- The standard for construct validity: convergent (high with related measures) and divergent (low with unrelated).
- It usually gives stronger correlations than change-based methods.
- Its assumptions are those of any correlation ([Topic 1](#)): linearity, a coefficient matched to the data, adequate range.



### KEY INSIGHT

A cross-sectional correlation shows the instrument properties at one moment. It does not show that it detects change over time.

**Supports:** cross-sectional construct validity (a between-patient property).



## Interpreting a correlation matrix

|           | New PROM                           | Related A                          | Related B                          | Unrelated                          |
|-----------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
| New PROM  | <b>1.00</b>                        | <b>0.72</b><br>95% CI [0.65, 0.78] | <b>0.65</b><br>95% CI [0.56, 0.72] | <b>0.18</b><br>95% CI [0.04, 0.31] |
| Related A | <b>0.72</b><br>95% CI [0.65, 0.78] | <b>1.00</b>                        | <b>0.58</b><br>95% CI [0.48, 0.67] | <b>0.22</b><br>95% CI [0.08, 0.35] |
| Related B | <b>0.65</b><br>95% CI [0.56, 0.72] | <b>0.58</b><br>95% CI [0.48, 0.67] | <b>1.00</b>                        | <b>0.20</b><br>95% CI [0.06, 0.33] |
| Unrelated | <b>0.18</b><br>95% CI [0.04, 0.31] | <b>0.22</b><br>95% CI [0.08, 0.35] | <b>0.20</b><br>95% CI [0.06, 0.33] | <b>1.00</b>                        |

### How to interpret it

- Set expectations first: related measures should have high correlation coefficients; unrelated measures should have low correlation coefficients.
- Report each correlation coefficient with a 95% confidence interval, and judge the size, not the statistical significance (benchmarks: [Topic 4](#)).
- The entire matrix shows the inference validity at a glance.

### REUSABLE CONCLUSION

As hypothesized, [New PROM] correlated strongly with [related measure A] ( $r = .72$ , 95% CI 0.65, 0.78) and had a weak relationship with [unrelated measure] ( $r = .18$ , 95% CI 0.04, 0.31), supporting convergent and divergent validity.

A new PRO converges with related measures and diverges from an unrelated one.



# A worked example

- New PROM vs an established related measure:  $r = 0.72$  (convergent, as expected).
- New PROM vs an unrelated measure:  $r = 0.18$  (divergent, as expected).
- Each coefficient is reported with a 95% confidence interval and judged by its size.



### IN PRACTICE

Each coefficient is checked against a pre-specified hypothesis. Together they build the cross-sectional construct-validity argument.

### R • RUNS ON THE EXAMPLE DATASET

```
# CDISC BDS is long: pivot to wide first
wide <- adcorr %>%
  filter(AVISIT == "Baseline") %>%
  pivot_wider(id_cols = USUBJID,
              names_from = PARAMCD,
              values_from = AVAL)

ct <- cor.test(wide$TARGTOT, wide$RELATEDA)
ct$estimate    # r = 0.73
ct$conf.int    # 95% CI 0.66 to 0.79
```

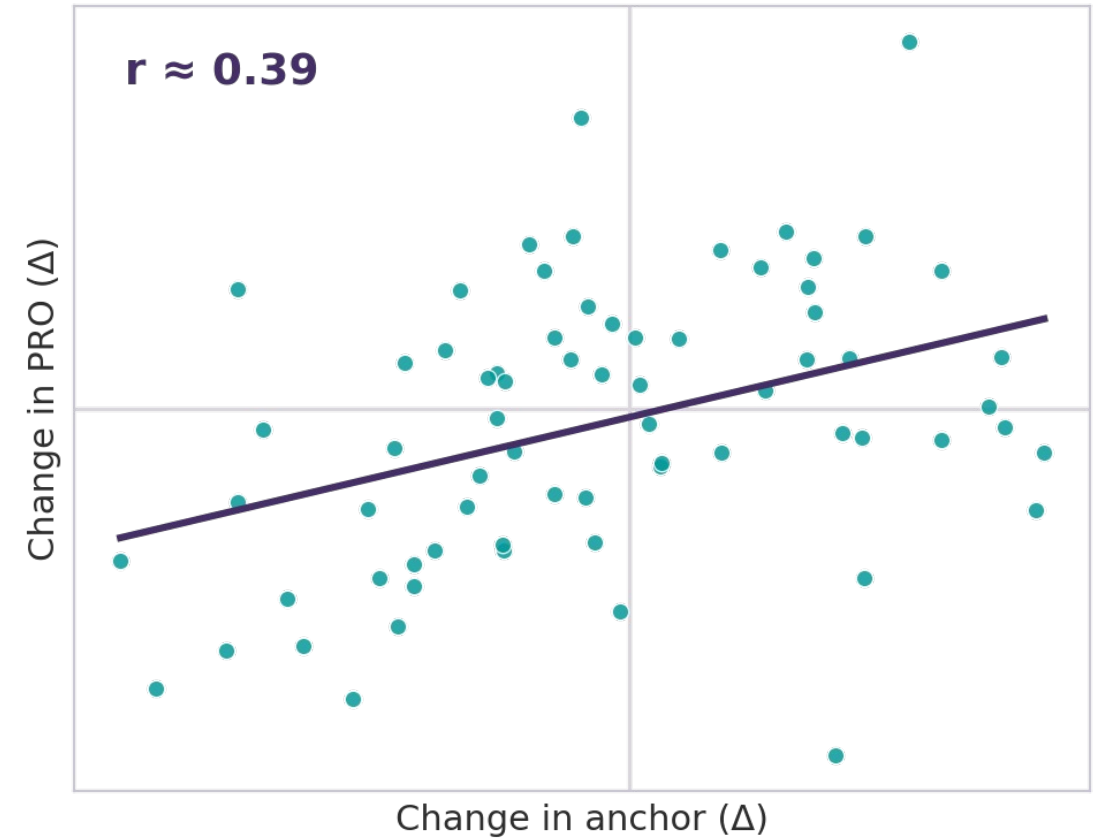


## Correlating change over time

A change-score correlation assesses whether patients who change on one measure also change on another. It is computed on change scores and reflects responsiveness, the longitudinal aspect of validity.

$$\Delta PRO = PRO_1 - PRO_0 \quad \Delta Anchor = Anchor_1 - Anchor_0 \quad \text{then} \\ r(\Delta PRO, \Delta Anchor)$$

- Change scores result from subtracting two measurements, so their errors compound (a change correlation is attenuated by the reliability of the change score itself; [reliability](#)): this is why change correlations are typically lower than cross-sectional ones (Cronbach & Furby, 1970).
- Use Pearson or Spearman correlation coefficient, matched to the distribution of the change scores ([Topic 3](#)).
- An anchor is an external measure used to interpret change (e.g. PGIS, CGIS, [anchors](#)); its quality drives the result. If the chosen anchor is not robust (its use has not been supported by previous evidence) or is used incorrectly, the resulting outcome may be difficult to interpret. Validity of anchors can be established qualitatively as there is no internal structure to be assessed. However larger packages of qualitative and quantitative checks can support anchor validity.



*Change in a PRO against change in an anchor, one point per patient.*

Thresholds for an adequate anchor correlation: Hays (2005): 0.37; Guyatt (2000) 0.5; acceptable: 0.3



## When it works and when it doesn't

### USE WHEN

- ✓ Changes in two measures may move together over time.
- ✓ Both the COA and the anchor are measured at baseline and follow-up (not anchors of change collected at follow up only).
- ✓ The anchor is theoretically related, interpretable, and pre-specified.
- ✓ There is enough variability in change among patients.

### AVOID WHEN

- ⚠ Data from only one timepoint is available.
- ⚠ Most patients stay stable, leaving little true change to correlate.
- ⚠ The anchor is weak (no previously validation evidence), poorly selected, not matched to the same or a related concept, or hard to interpret.
- ⚠ Change is dominated by error, floor/ceiling effects, or timepoint mismatch (different timepoints used).



## Interpret it as hypothesis-driven evidence

Interpret a change correlation as pre-specified evidence of expected change, not as stand-alone proof of responsiveness or validity. *It supports responsiveness but can't confirm it alone*: the value also depends on the anchor, and how much true change occurred.

- Consider the scoring direction: improvement can give a positive or negative correlation depending on coding or scoring direction ([scoring direction](#)).
- Do not treat a weak correlation as automatic proof of non-responsiveness.
- Judge against the pre-specified hypothesis and context of use, not statistical significance.



### WATCH OUT

If the strength of the correlation is not aligned with expectations, it does not necessarily mean there is an issue with the COA. It may reflect a weak anchor, little true change, timing mismatch, a restricted range, or measurement error in the change scores.

**Further pitfalls** (measurement error, floor/ceiling, regression to the mean) are covered in [Topic 5](#).



### An example

- Objective: Compare the correlation score coefficient against the planned threshold and report with a 95% CI.
- Result obtained: Change in a PRO vs change on a global rating (CGIS):  $r = 0.45$  (moderate).
- Conclusion: This result is consistent with the pre-specified responsiveness hypothesis.

#### REUSABLE CONCLUSION

*Change in [PRO] correlated moderately with change in [anchor] ( $r = 0.45$ , 95% CI ...), consistent with the pre-specified responsiveness hypothesis.*

#### R • RUNS ON THE EXAMPLE DATASET

```
# BASE and CHG are pre-computed in BDS
d <- adcorr %>%
  filter(AVISIT == "Week 12") %>%
  pivot_wider(id_cols = USUBJID,
              names_from = PARAMCD,
              values_from = c(AVAL, CHG))

# change in the PRO against a change anchor
cor.test(d$CHG_TARGTOT, d$AVAL_PGIC) # 0.40
```



## When data are nested, a correlation isn't one number

When the same people are measured repeatedly, their observations are not independent. A single correlation then blends two different associations: differences between people, and changes within a person over time.



### KEY INSIGHT

A correlation from nested data is an answer to an unspecified question, until you say which level it describes.  
One correlation, two levels: always specify if you are measuring differences between individuals or variation over time.

- Failing to account for the hierarchical nature of the data does not just lose efficiency: it can answer the wrong clinical question.

### NESTING: VISITS WITHIN PATIENTS



*Patient* → *four monthly visits*



### IN PRACTICE

60 patients × 4 visits = 240 rows, but only 60 independent units.



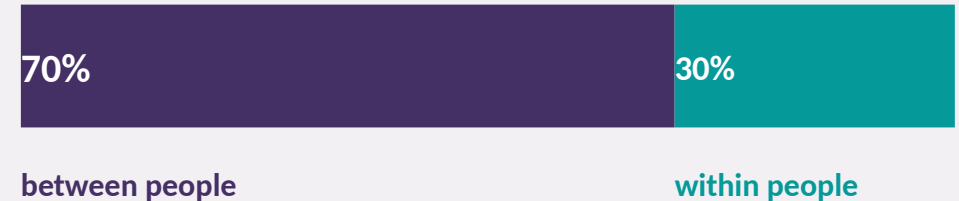
## Nested data splits variance into two parts

Repeated observations from one person share variance that comes from stable individual traits. The ICC is the share of total variance that lies between people; the rest lies within people, throughout time points.

- High ICC: most information is between people, so a simple total correlation is dominated by the between-subject association.
- Low ICC: within-person variation is large, so the total correlation can hide a meaningful within-subject signal.
- Very high ICC (0.8–0.9 or more): small within-person variation means that within-subject correlation rests on low signal.

*This volume presents the ICC generically. In practice, several ICC variants exist (ICC(2,1), ICC(3,1), Shrout & Fleiss' 1979), which use different variance components and can give different values for the same data.*

Example: ICC = 0.70



### IN PRACTICE

Interpret the ICC as a diagnostic step, before proceeding with the analysis: it tells you how much of any correlation will come from each level, and so which method and which level matter (Curran & Bauer, 2011).

*Because the variance from nested data splits between and within, so does any correlation computed from the data.*



## WITHIN & BETWEEN SUBJECTS • THE DISTINCTION

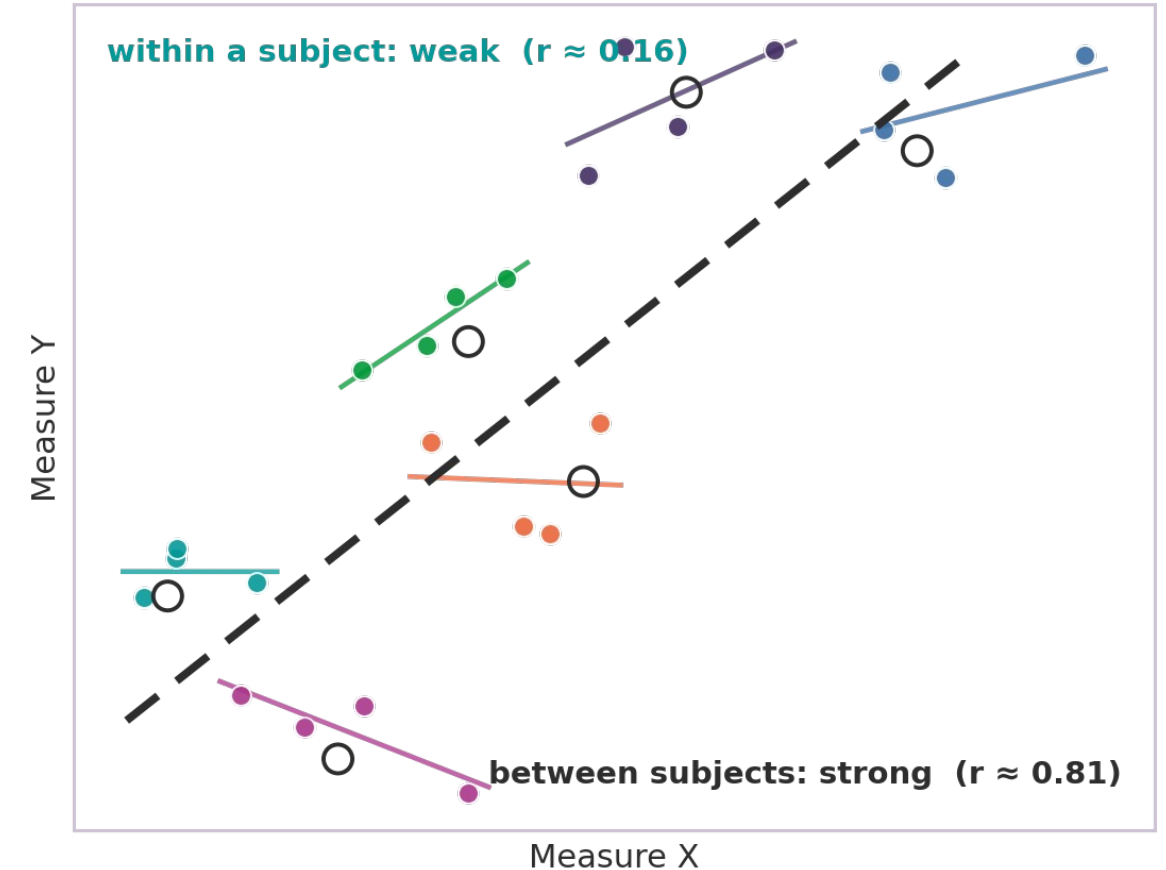
# One dataset, three correlations

- **Total** Across all rows, ignoring structure. Conflates both levels; do not interpret it on its own.
- **Between** Based on the overall patient averages.  
Do patients who score higher on X also tend to score higher on Y?
- **Within** Among each patient's deviations from their own mean.  
When a patient rises above their usual X, does Y rise too?



### WATCH OUT

These three types of correlations can differ in size, and sometimes in sign ([Simpson's paradox](#); Simpson, 1951).



Six patients with four visits each.

Subject means (open circles) rise together; most patient's own cloud are barely associated



## Two questions, two kinds of construct validity



When the question is whether an instrument detects change **within a patient**, the within-subject correlation, rather than the between-subject one, is the evidence for validity.

### Between-subject

→ *cross-sectional construct validity*

Patients with worse disease report worse scores. **Evidence:** convergent and divergent correlations among patients.

### Within-subject

→ *longitudinal validity & responsiveness*

When a patient's disease worsens, their score worsens with it. **Evidence:** correlate within-patient change in the score with change in an anchor over time.

**On regulation:** FDA and EMA look for evidence *of an instrument's ability to detect change, not only the ability to discriminate between patients*. In theory, strong cross-sectional correlation should imply this; in practice, reviewers still expect to see the longitudinal correlation evidence.



## Three ways to separate the levels

With the ICC understood ([ICC](#)), choose how to separate the within- and between-subject associations:



### Person-mean centering

Subtract each person's mean; any software.  
Within and between, separately.

*Use as the transparent default.*



### rmcorr

One pooled within-subject estimate with a CI (R  
rmcorr). Assumes a common slope.

*Use for a single within answer.*



### Mixed-effects models

Each patient gets their own slope; can model  
AR(1) over time (lme4 / nlme).

*Use when trajectories differ or data are  
autocorrelated (e.g. daily pain).*

*Decide which level answers your question before you look at the data.*

*The interpretation patterns are in the appendix.*



# Person-mean centering

## How it works

For each person, subtract their own mean from each observation. Correlate the centered scores for the within-subject association; correlate the person means for the between-subject one.

- Runs in any software, with no special package.
- Gives within and between as separate, transparent numbers.
- Needs three to four occasions per person; occasion-level missing data lowers precision.



## IN PRACTICE

The most transparent approach (Curran & Bauer, 2011).

## R • RUNS ON THE EXAMPLE DATASET

```
library(dplyr)

d <- adcorr %>%
  filter(PARAMCD %in% c("FINTOX", "SYMBUR")) %>%
  pivot_wider(id_cols = c(USUBJID, AVISITN),
              names_from = PARAMCD,
              values_from = AVAL) %>%
  group_by(USUBJID) %>%
  mutate(fin_w = FINTOX - mean(FINTOX),
         sym_w = SYMBUR - mean(SYMBUR)) %>%
  ungroup()

cor(d$FINTOX, d$SYMBUR) # naive 0.54
cor(d$fin_w, d$sym_w) # within 0.23
```



## Repeated-measures correlation (rmcorr)

### How it works

Introduced by Bland & Altman (1995) and formalized as rmcorr (Bakdash & Marusich, 2017). It returns one within-subject correlation pooled across everyone, adjusting for non-independence by ANCOVA.

- Gives a single within-subject  $r$  with a confidence interval.
- Good when you want one clear within-subject answer.



### WATCH OUT

rmcorr assumes a common within-subject slope. It does not estimate individual slopes and cannot model autocorrelation (no AR(1)). For those, use a mixed model.

### R • RUNS ON THE EXAMPLE DATASET

```
library(rmcorr)

# one within-subject correlation, pooled
# among patients, adjusting for
# non-independence (ANCOVA)
fit <- rmcorr(participant = USUBJID,
              measure1 = FINTOX,
              measure2 = SYMBUR,
              dataset = as.data.frame(d))
fit$r      # 0.23      fit$CI  # 0.16 to 0.31
```



# Mixed-effects models (lme4 / nlme)

## How it works

A mixed model gives each patient their own intercept and slope, so patients can follow different trajectories over time. It can also model how residuals are correlated across closely spaced visits, for example an AR(1) structure.

- Different trajectories: use it when patients' within-subject slopes genuinely differ.
- Autocorrelation: use it when closely spaced assessments, such as daily pain diaries, leave AR(1) residuals.
- Sample size: random slopes and AR(1) add parameters, so they need enough patients and enough occasions each; with few of either, simplify the model.



## IN PRACTICE

Reach for this when rmcrr's single-slope, no-AR(1) assumptions are too restrictive.

## R · RUNS ON THE EXAMPLE DATASET

```
library(nlme)

# random intercept per patient, AR(1) on
# the day-to-day residuals
ar1 <- corAR1(form = ~ AVISITN | USUBJID)
m <- lme(DIARY2 ~ DIARY1,
        random = ~ 1 | USUBJID,
        correlation = ar1, data = d)

cor(d$DIARY1, d$DIARY2) # naive 0.24
# within-person 0.54    AR(1) phi 0.42
```



## READING THE OUTPUT

The fixed slope for x is the within-subject association, interpret it on the same scale as r. The random-effects variances split total variability into between- and within-subject parts. The corAR1 parameter shows how strongly residuals are autocorrelated over time.



## Financial toxicity and symptom burden

**Synthetic study, realistic parameters.** 38 oncology patients completed a financial-toxicity PRO and a symptom-burden measure at four monthly visits (152 observations). Levels separated by person-mean centering.

| Level              | r           | Interpretation                                                                                                                           |
|--------------------|-------------|------------------------------------------------------------------------------------------------------------------------------------------|
| Total (unadjusted) | <b>0.58</b> | Moderate, but which level does it reflect?                                                                                               |
| Between-subject    | <b>0.71</b> | Patients with persistently higher financial toxicity report higher symptom burden.                                                       |
| Within-subject     | <b>0.24</b> | Month-to-month changes in financial toxicity are only weakly linked to concurrent symptom changes ( <a href="#">Simpson's paradox</a> ). |



### KEY INSIGHT

Diverging results: the total (0.58) hides a strong between-subject link (0.71) and a weak within-subject one (0.24). For a trial of a financial intervention, the within-subject answer is the important one.



## Checklist and takeaways

### BEFORE YOU REPORT A CORRELATION FROM REPEATED MEASURES



State which correlation you mean: total, between, or within.



Account for the non-independence of repeated observations.



Use three to four occasions per subject for within-subject estimates.



Keep cross-sectional validity and longitudinal responsiveness distinct.



Check whether within and between differ in size or reverse in sign.

1

A correlation from nested data is a mixture of within and between. Decompose before interpreting.

2

The two levels answer different questions and can even reverse in sign (Simpson's paradox).

3

For individual-level responsiveness you need within-subject evidence; strong cross-sectional validity supports but doesn't substitute.

TOPIC 2 - APPENDIX

# Within- and between-subject correlation

Depth, technical detail, and references



## Application conditions and assumptions



**Sufficient occasions per subject.** Within-subject estimates from only two timepoints are highly unstable. Three to four occasions are a practical minimum; power rises with both people and occasions (Bakdash & Marusich, 2017).



**Adequate within-person variability.** If a person's scores barely move across occasions, there is no within-subject signal to detect. Distinguish a genuinely stable construct from measurement insensitivity.



**Non-independence must not be ignored.** Treating repeated observations as independent is an error, not a conservative choice: standard errors are wrong and confidence intervals misleadingly narrow.



**Timing and spacing of assessments.** Occasions must suit the construct and question. Too close may reflect carry-over; too far apart may capture change that obscures the association of interest.



**Missing-data handling.** Match the strategy to the method: FIML (Full Information Maximum Likelihood) for model-based approaches, careful occasion-level handling for person-mean centering. Pre-specify.



## Interpretation patterns

There are no universal magnitude thresholds for these decompositions. The question is whether within and between differ meaningfully from each other and from the total, and what that implies for the validity claim.

This framework applies only to repeated-measures data, a single cross-sectional correlation has no within-subject component to isolate.

### Convergence

Within  $\approx$  between  $\approx$  total. The association is consistent across levels. Standard cross-sectional validity evidence is likely sufficient.

### Divergence

Between substantially larger than within (e.g. 0.71 vs 0.24). Captures stable differences well but is less sensitive to change; cross-sectional evidence overstates responsiveness.

### Reversal

Within and between differ in sign: a Simpson's paradox. The aggregate association is entirely misleading about individual-level dynamics.

*Generic pitfalls (treating repeated observations as independent, ecological-fallacy reasoning) are covered in [Topic 5: Challenges and solutions](#).*



## References

- Slide 22** Cronbach, L. J., & Furby, L. (1970). How we should measure “change”: Or should we? *Psychological Bulletin*, 74(1), 68-80;
- Hays, R. D., Farivar, S. S., & Liu, H. (2005). Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD*, 2(1), 63–67;
- Guyatt G. H. (2000). Making sense of quality-of-life data. *Medical care*, 38(9 Suppl), II175–II179.
- Slide 30** Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annu Rev Psychol*, 62, 583-619.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*, 86(2), 420-428;
- Slide 29** Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *J R Stat Soc B*, 13(2), 238-241.
- Slide 30** FDA (2009). PRO Measures: Use in Medical Product Development to Support Labeling Claims;
- FDA (2025). Patient-Focused Drug Development: Fit-for-Purpose COAs;
- EMA (2016). Appendix 2, anticancer medicinal products: PRO measures in oncology.
- Slide 34** Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annu Rev Psychol*, 62, 583-619.
- Slide 33** Bland, J. M., & Altman, D. G. (1995). Calculating correlation coefficients with repeated observations. *BMJ*, 310, 446;
- Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Front Psychol*, 8, 456.

### Volume-wide further reading:

- Bates, D., et al. (2015). Fitting linear mixed-effects models using lme4. *J Stat Softw*, 67(1), 1-48;
- Pinheiro, J. C., & Bates, D. M. (2000). Mixed-Effects Models in S and S-PLUS. *Springer*.



TOPIC 3

# Types of correlation

Matching the coefficient to the data and the question



## Match the coefficient to the data

The right coefficient depends on how each variable is measured. A correlation is a model-based summary; the wrong choice attenuates or distorts the association.

This framework assumes every patient answered the same fixed item set.

|            | Continuous                | Ordinal                 | Binary                    |
|------------|---------------------------|-------------------------|---------------------------|
| Continuous | Pearson / Spearman        | Polyserial / Spearman   | Point-biserial / Biserial |
| Ordinal    | Polyserial                | Polychoric / Spearman   | Polychoric <sup>1</sup>   |
| Binary     | Point-biserial / Biserial | Polychoric <sup>1</sup> | Tetrachoric / Phi         |

<sup>1</sup> a binary variable treated as a two-level ordinal (e.g. via the *psych* package in R).

### IN OTHER WORDS

Different correlation coefficients answer different measurement questions:

- Pearson summarizes linear observed-score association,
- Spearman summarizes monotonic rank-based association,
- polychoric/polyserial correlations estimate latent associations under threshold assumptions

### Interpreting the matrix

Find your two data types on the edges; the cell names the coefficient. Continuous–continuous is the familiar Pearson / Spearman case; the other cells handle ordinal and binary data.



### IN PRACTICE

In COA work, many variables are ordinal or binary observations of continuous health concepts, so the latent-variable coefficients (polychoric, polyserial, biserial) matter. For agreement on continuous repeated measures, use the [ICC](#).



# Pearson correlation (r)



## WHAT IT IS

A standardized measure of the linear association between two continuous variables.

## UNDER THE HOOD

Pearson divides the covariance of the two variables by the product of their standard deviations. That standardisation makes it scale-free and bounded by -1 to +1.

## EXAMPLE COA USES

- Convergent validity between numeric or visual-analogue scales, or domain / total scores.
- Associations between change scores (responsiveness).

## KEY LIMITS

- Assumes interval-level measurement and linearity.
- Sensitive to outliers, non-linearity, and measurement error.
- Attenuates associations when variables are ordinal.

## R • RUNS ON THE EXAMPLE DATASET

```
# CDISC BDS is long: pivot to wide first
library(dplyr); library(tidyr)
wide <- adcorr %>%
  filter(AVISIT == "Baseline") %>%
  pivot_wider(id_cols = USUBJID,
              names_from = PARAMCD, values_from = AVAL)

cor(wide$TARGTOT, wide$RELATEDA)      # 0.73
```



## Spearman rank correlation ( $\rho$ )



### WHAT IT IS

A correlation computed on the ranked values of two variables, not their raw scores. It's a non-parametric statistic.

### UNDER THE HOOD

Spearman replaces each value with its rank, then applies the Pearson formula to those ranks. So it captures any monotonic relationship, not just a straight line.

### EXAMPLE COA USES

- Item-item relationships for ordinal or non-normal variables.
- Ordinal anchors (PGIS, CGIS) when a latent model (see next) is not needed. The decision not to use a latent variable needs to be justified.
- Robustness checks alongside Pearson.

### KEY LIMITS

- No variance-based interpretation, unlike Pearson.
- Still attenuates when responses collapse into a few broad categories.
- Kendall's  $\tau$  is an alternative for small samples or many ties ([reliability](#)).

### R • RUNS ON THE EXAMPLE DATASET

```
# same wide layout as Pearson
cor(wide$ITEM02, wide$ITEM03,
    method = "spearman")           # 0.62

# Kendall alternative for many ties:
cor(wide$ITEM02, wide$ITEM03,
    method = "kendall")           # 0.55
```



# Polyserial correlation



## WHAT IT IS

A correlation method for one continuous and one ordinal variable, assuming the ordinal variable reflects an underlying continuous latent variable with thresholds.

## UNDER THE HOOD

Polyserial correlation corresponds to the correlation between the observed continuous variable and a latent continuous variable underlying the ordinal measure (not the correlation between observed scores).

## EXAMPLE COA USES

- Item-domain or item-total relationships (ordinal item vs continuous score, [corrected item-total correlation](#))
- Ordinal anchors (e.g., PGIS, PGIC) vs continuous endpoints.
- Convergent validity where constructs are assessed using mixed measurement scales

## KEY LIMITS

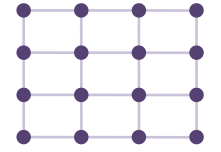
- Relies on latent normality and threshold assumptions for the ordinal variable
- Sensitive to sparse or highly skewed response categories
- More complex to estimate than Pearson or Spearman correlations (requires additional assumptions and specialized estimation methods)
- Reflects a model-based association, not a directly observed relationship

## R • RUNS ON THE EXAMPLE DATASET

```
library(polycor)

# ordinal PGIS against continuous TARGTOT
polyserial(wide$TARGTOT, wide$PGIS) # 0.72

#Compare vs Pearson
cor(wide$TARGTOT, wide$PGIS) # 0.68
```



*ordinal × ordinal*

# Polychoric correlation

## WHAT IT IS

A correlation method for two ordinal variables, assuming each reflects an underlying continuous latent variable with thresholds.

## UNDER THE HOOD

Correlation between the latent continuous variables (not the correlation between observed categories).

## EXAMPLE COA USES

- Item-item relationships for ordinal response scales
- Used in the underlying correlation matrix in EFA and CFA (must be specified in code)

## KEY LIMITS

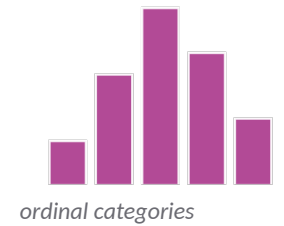
- Relies on latent normality and threshold assumptions
- Sensitive to sparse categories or skewed distributions
- More complex to estimate and interpret than Pearson/Spearman (requires additional assumptions and specialized estimation methods)
- Reflects a model-based association, not directly observed relationships

## R • RUNS ON THE EXAMPLE DATASET

```
library(polycor)

# two ordinal items, same wide layout
polychor(wide$ITEM02, wide$ITEM03) # 0.71

# full item matrix (psych package):
psych::polychoric(
  wide[, c("ITEM02", "ITEM03")])$rho
```



## Coefficients for ordinal and binary data

| Coefficient    | Variables                          | Assumes Latent variable | Typical COA use                                                                                         |
|----------------|------------------------------------|-------------------------|---------------------------------------------------------------------------------------------------------|
| Point-biserial | binary (true) × continuous         | No (Pearson case)       | group or sex × score; item correct × total                                                              |
| Biserial       | binary (dichotomized) × continuous | Yes                     | an artificially split continuous variable × score (e.g. a median split. Note: never use a median split) |
| Polychoric     | ordinal × ordinal                  | Yes (both)              | Likert item × item; input to factor analysis                                                            |
| Polyserial     | ordinal × continuous               | Yes (ordinal)           | ordinal item × total or continuous score                                                                |
| Tetrachoric    | binary × binary                    | Yes (both)              | two dichotomous items                                                                                   |



### KEY INSIGHT

The latent-model coefficients estimate the correlation between the assumed underlying continuous variables, so they recover what Pearson attenuates on coarse categories. This is central for ordinal Likert items.

### R • RUNS ON THE EXAMPLE DATASET

```
library(psych)
cor(wide$TARGTOT, wide$SEVGRP)      # 0.46
cor(wide$TARGTOT, wide$DICHOTOM)    # 0.37
biserial(wide$TARGTOT, wide$DICHOTOM) # 0.48
```



## The use and abuse of the ICC

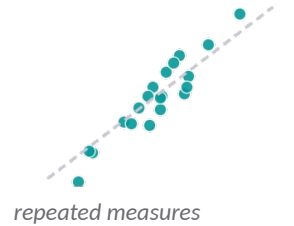
The ICC is a correlation for repeated measurements, built from the same standardized-covariance idea but answering a different question.

**Grand mean:** the overall average across all people and all their measurements (a group-level reference, not each person's own mean).

**The ICC asks:** when one of a person's measurements sits above (or below) the grand mean, do their other measurements too? It partitions the total variance into between-person and within-person parts.

**General definition:** the ICC quantifies the proportion of total variance attributable to differences between individuals. Equivalently, it describes the extent to which repeated measurements or ratings from the same person remain similar.

*Therefore it suits test-retest, inter-rater and intra-rater reliability, and nested data (Topic 2).*



### STATE ITS FORM, OR IT IS AMBIGUOUS

**Model:** one-way vs two-way (are raters or occasions random?)

**Definition:** absolute agreement vs consistency.

**Unit:** a single measure vs the average of several.

*ICC = 0.87" alone cannot be interpreted. Shrout & Fleiss (1979); McGraw & Wong (1996).*

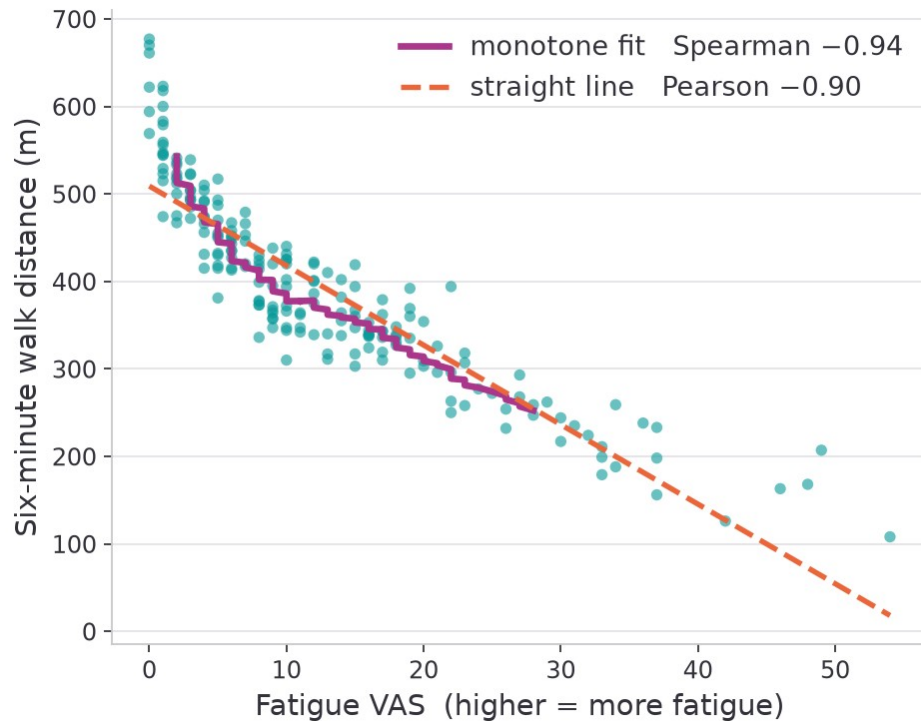
**REUSABLE CONCLUSION:** "Test-retest reliability was good: ICC = 0.87 (95% CI 0.83, 0.90), two-way random-effects model, absolute agreement, single measures.

### R • RUNS ON THE EXAMPLE DATASET

```
# wide: rows = subjects, one column per visit
retest <- adcorr %>%
  filter(PARAMCD == "TARGTOT", AVISITN < 12) %>%
  pivot_wider(id_cols = USUBJID,
              names_from = AVISIT, values_from = AVAL)
psych::ICC(retest[, -1]) # ICC(2,1) 0.87
```

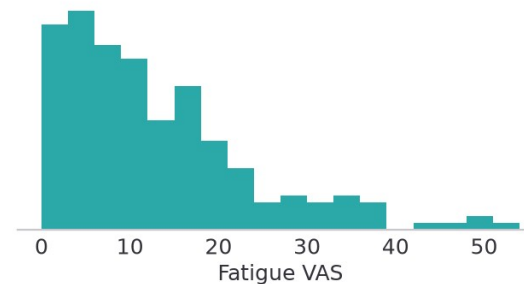


## Curved and skewed: Pearson underestimates the relationship



### THE DATASET

200 patients: a fatigue VAS against six-minute walk distance. More fatigue goes with less walking. The relationship is strong but curved, and the VAS is right-skewed.



### IN PRACTICE

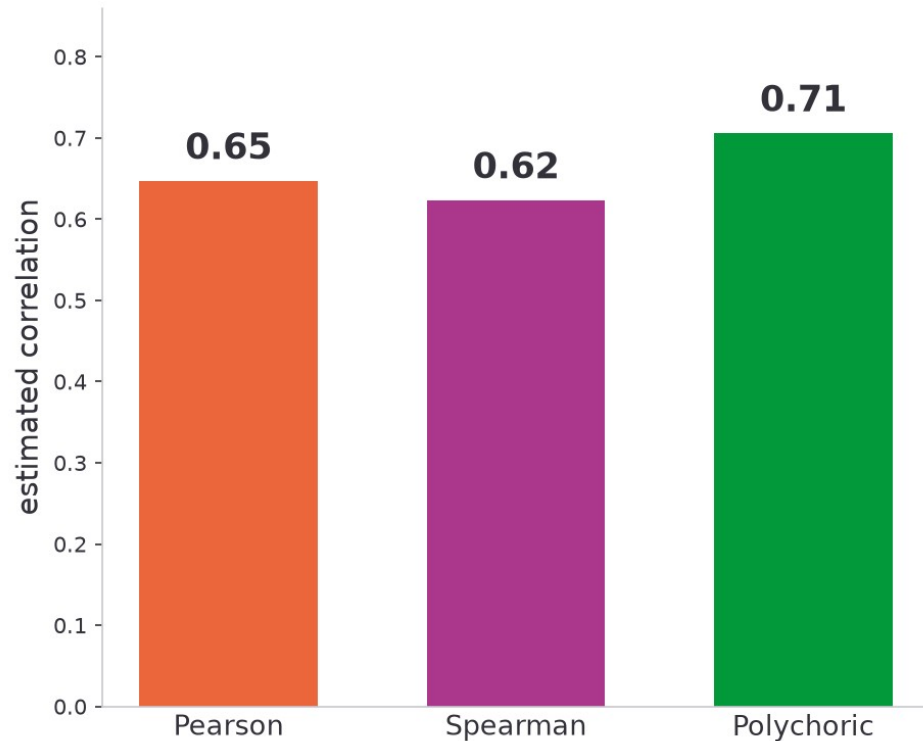
With a monotonic but non-linear relationship, report Spearman; Pearson is pulled down by the curvature, not by a weak association.

*Note: FDA often asks for Spearman correlations to be reported.*

*Pearson  $r = -0.90$  already reads as strong, but Spearman  $\rho = -0.94$  shows the monotonic association is stronger still.*



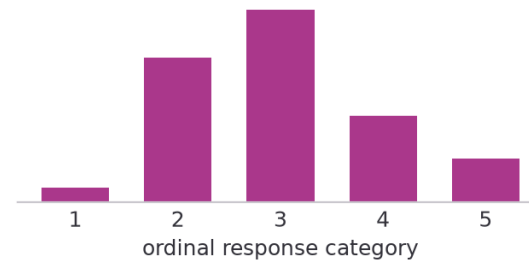
## Coarse ordinal data: polychoric recovers the signal



Pearson 0.65 and Spearman 0.62 attenuate the association. Polychoric recovers it, returning 0.71 against a true value of 0.71.

### THE DATASET

Two 5-point Likert items from the same domain,  $n = 200$ . Responses spread across all five categories, with the middle ones used most.



*Note: ordinal response coding, 1 to 5, is a convention to analyze the data, there is no inherent value to these numeric values.*



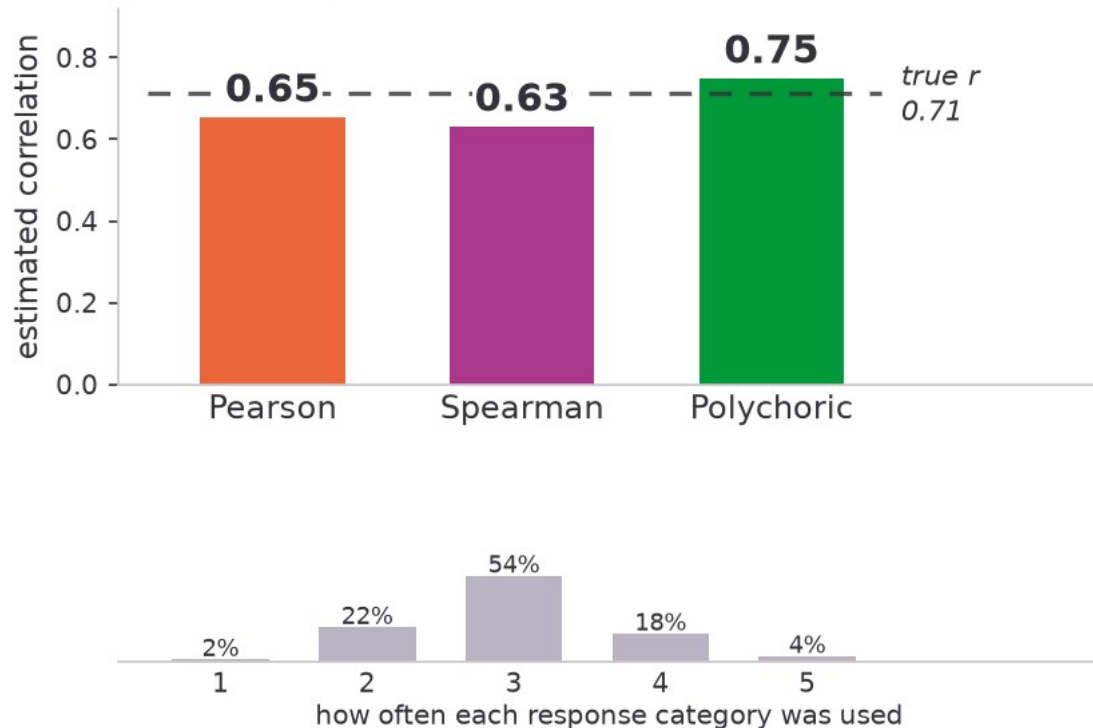
### IN PRACTICE

On few ordinal categories, Pearson and Spearman attenuate; polychoric estimates the correlation between the underlying constructs. It is the best default here, not an exact answer everywhere.

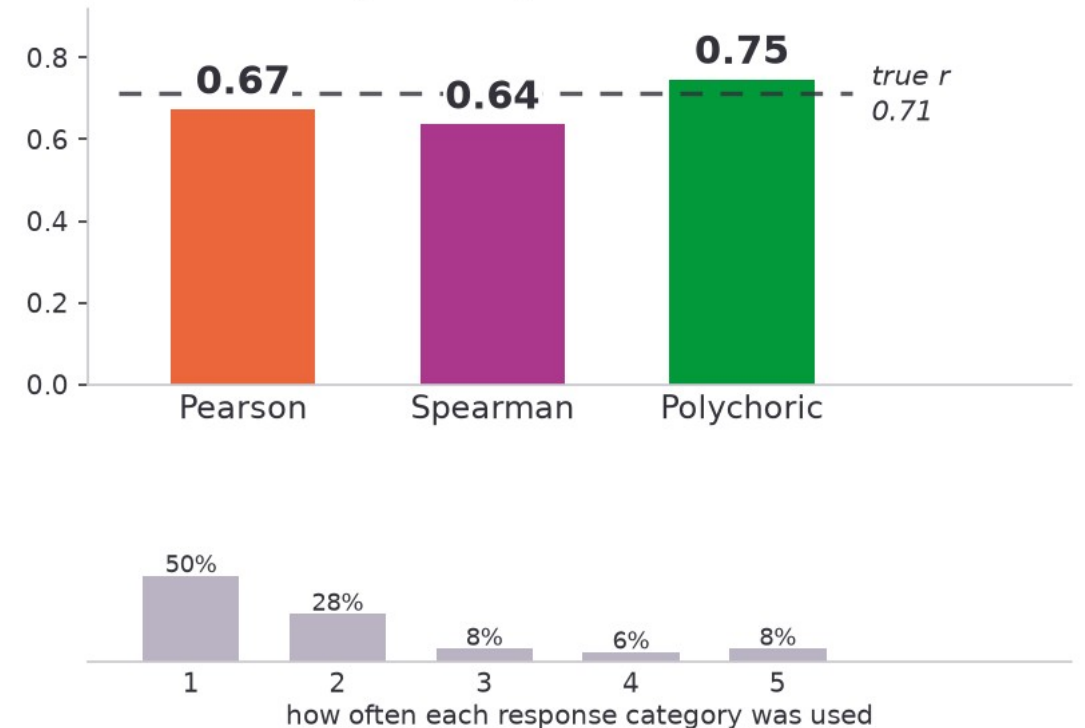


## When responses bunch up, polychoric still wins, but it starts to overshoot too

Responses bunched in the middle



Responses piled on the floor

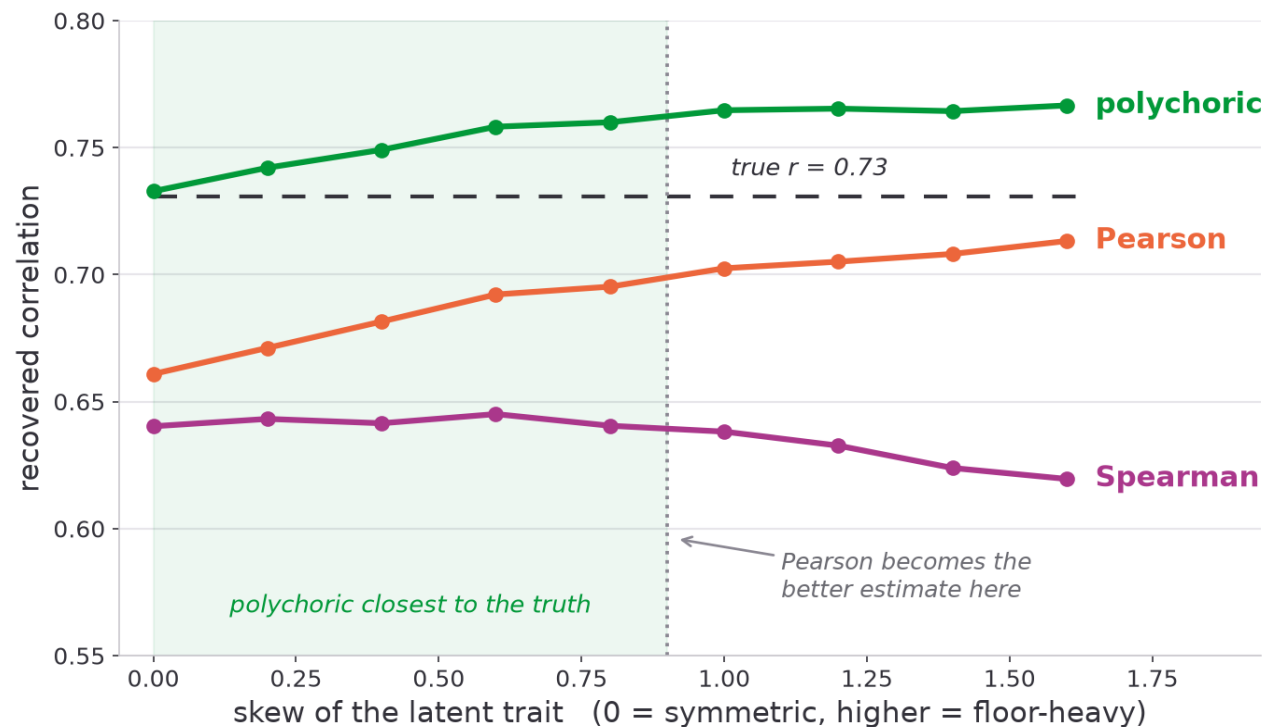


The same four items and the same latent truth as the previous slide, redrawn under two response distributions clinicians meet constantly. Polychoric stays the closest of the three, but both shapes push it above the true 0.71.



## Does the right coefficient always recover the truth?

Each coefficient was asked to recover a known underlying correlation ( $r \approx 0.73$ ) from 5-point items, as the latent trait shifts from symmetric to floor-heavy, the shape symptom data can often take in real patients.



$n = 200$  • 5-point items • true latent  $r \approx 0.73$  • mean of 500 simulations per skew level. Polychoric is the safest default, but no single coefficient is unconditionally exact.

### POLYCHORIC

Recovers the underlying correlation when responses are near-symmetric. Stays closest to it across mild skew. The only one of the three that targets the latent association.

### PEARSON

Attenuated when the latent trait is near-normal, but de-attenuates as the distribution skews. Past a skew of about 0.9 it becomes the closer estimate. Its error is not constant.

### SPEARMAN

Never recovers the latent correlation, sitting near 0.63 throughout. It answers a rank-based question, so if used to estimate an inter-item or latent  $r$  it will produce a conservative estimate.



## Five things to carry forward



### Match coefficient to data

Choose by scale, information content, and whether the variable is observed or latent.



### Pearson vs Spearman

Pearson for linear continuous; Spearman for monotonic, ordinal, or non-normal data.



### Recover the latent correlation

For ordinal and binary data, polychoric / polyserial / biserial / tetrachoric avoid attenuation. This matters for Likert items.



### State the ICC's form

The ICC is a correlation for repeated measures and agreement; name the model, definition, and unit.



### It can change the conclusion

The wrong coefficient matters most with few categories, skew, or latent constructs.



## References

**Slide 45** Spearman, C. (1904). The proof and measurement of association between two things. *Am J Psychol*, 15, 72-101.

**Slide 49** Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*, 86(2), 420-428;  
McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychol Methods*, 1(1), 30-46.

### Volume-wide further reading:

Revelle, W. (2024). psych: Procedures for Psychological, Psychometric, and Personality Research. R package (polychoric, ICC).

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443-460.



TOPIC 4

# Correlation outputs

Showing, estimating, and interpreting a correlation



## What a correlation report needs

Three things that should be presented together. A correlation coefficient on its own will lack important context.



### A picture

Plot the data before you trust the number.

*Scatter plots and related displays.*



### An estimate with a CI

Report  $r$  with a 95% confidence interval.

*Precision, not just a point value.*



### A benchmarked reading

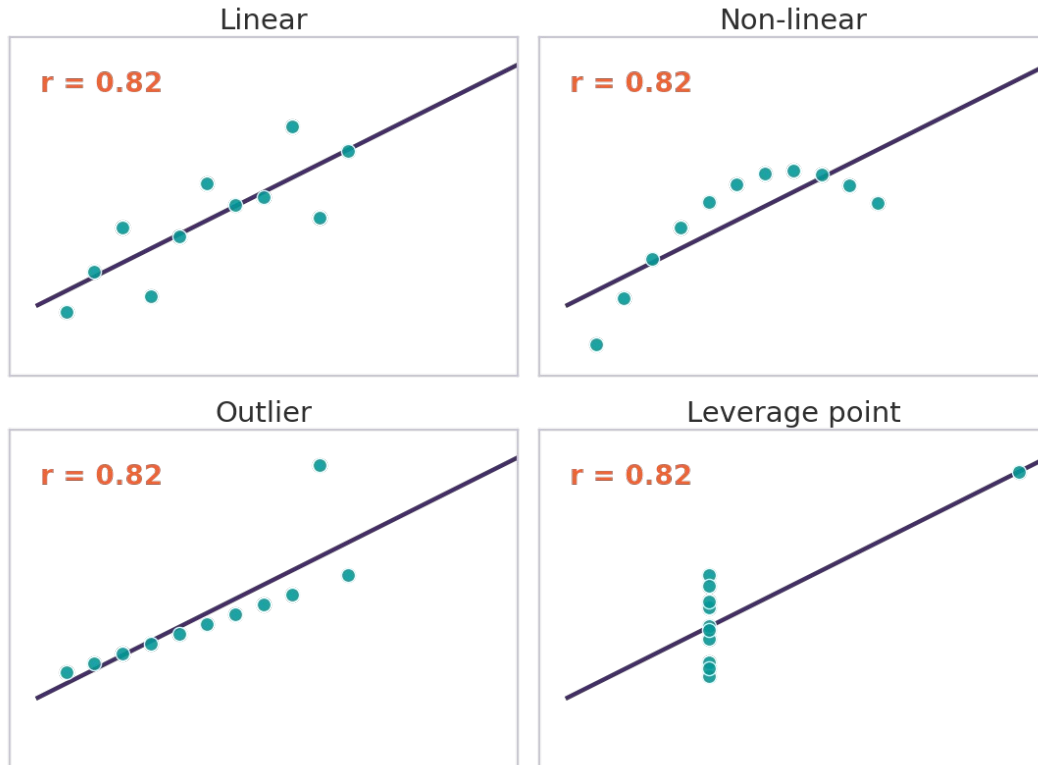
Judge the size against a pre-specified benchmark. ( $R^2$ )

*In the context of the measurement purpose.*

**Statistical significance is rarely the answer**; the size of the correlation and the precision around it are what's important to the interpretation.



## Always plot before you trust the number



*Anscombe's quartet: four datasets, all with  $r = 0.82$  (Anscombe, 1973).*

### A single $r$ can hide

- Non-linear relationships that  $r$  reads as weak.
- Outliers or leverage points that create or mask a correlation.
- Range restriction that shrinks  $r$ .
- Subgroups or clusters posing as one association.

**The next four slides:** the plots that help add context – scatter, correlation matrix, Bland–Altman, and spaghetti.

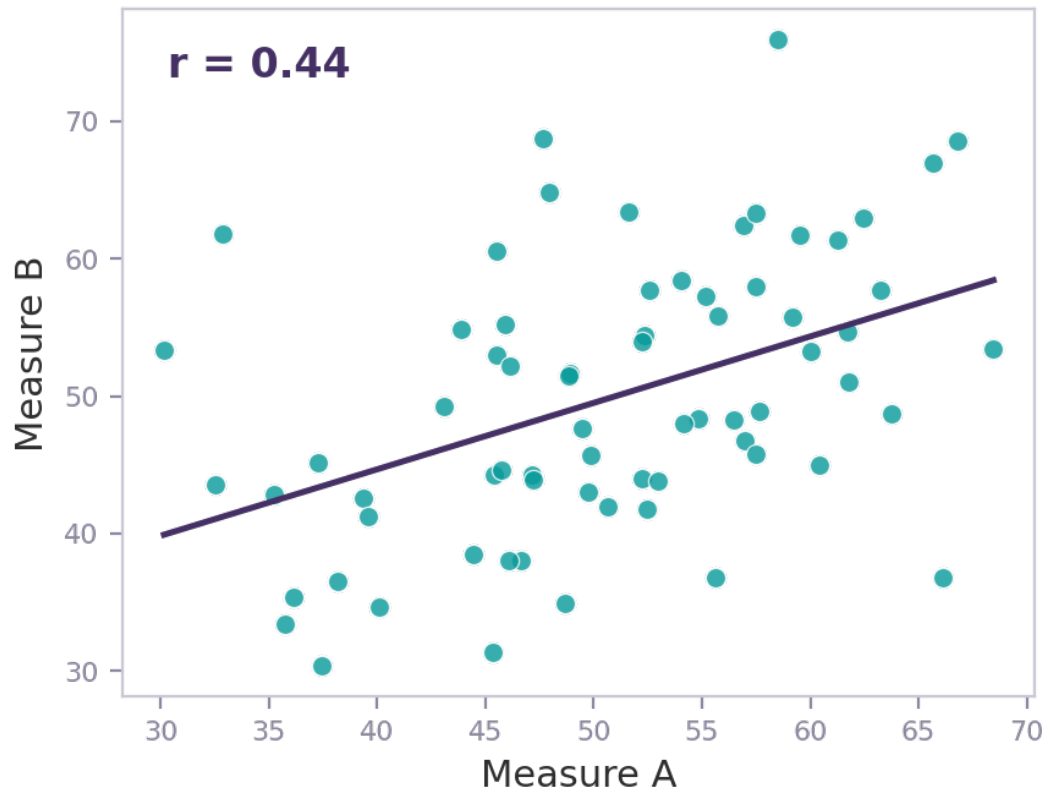


### KEY INSIGHT

The coefficient is a summary. The real evidence is the plot we made along the way.



## Scatter plot: the default for two continuous measures



### WHAT IT SHOWS

- The form of the relationship – linear, curved, or none – all at a glance.
- Outliers and leverage points, spread, and clustering.
- Simply put - whether a single  $r$  is a fair summary of the cloud.

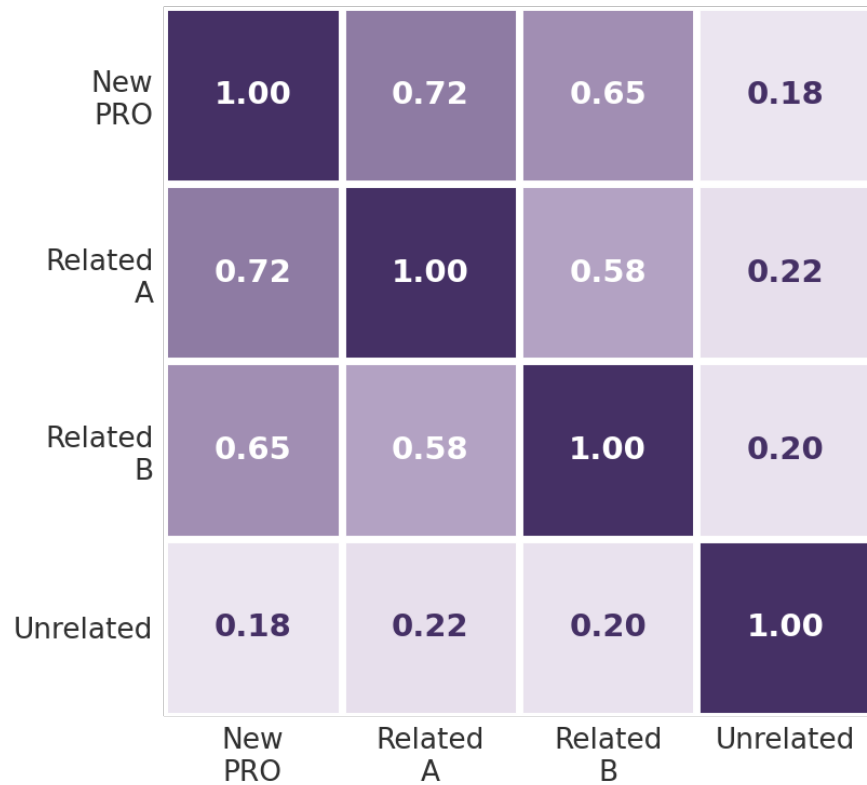
### R • RUNS ON THE EXAMPLE DATASET

```
library(ggplot2)

# wide: one row per subject
ggplot(wide, aes(TARGET, RELATEDA)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm")
```



## Correlation matrix: every pair at once



### WHAT IT SHOWS

- Every pairwise correlation in a set of items or scales, as a heat map.
- Convergent blocks (high) and divergent patterns (low) stand out at once.
- The usual first look at a multi-item or multi-domain COA.

### R • RUNS ON THE EXAMPLE DATASET

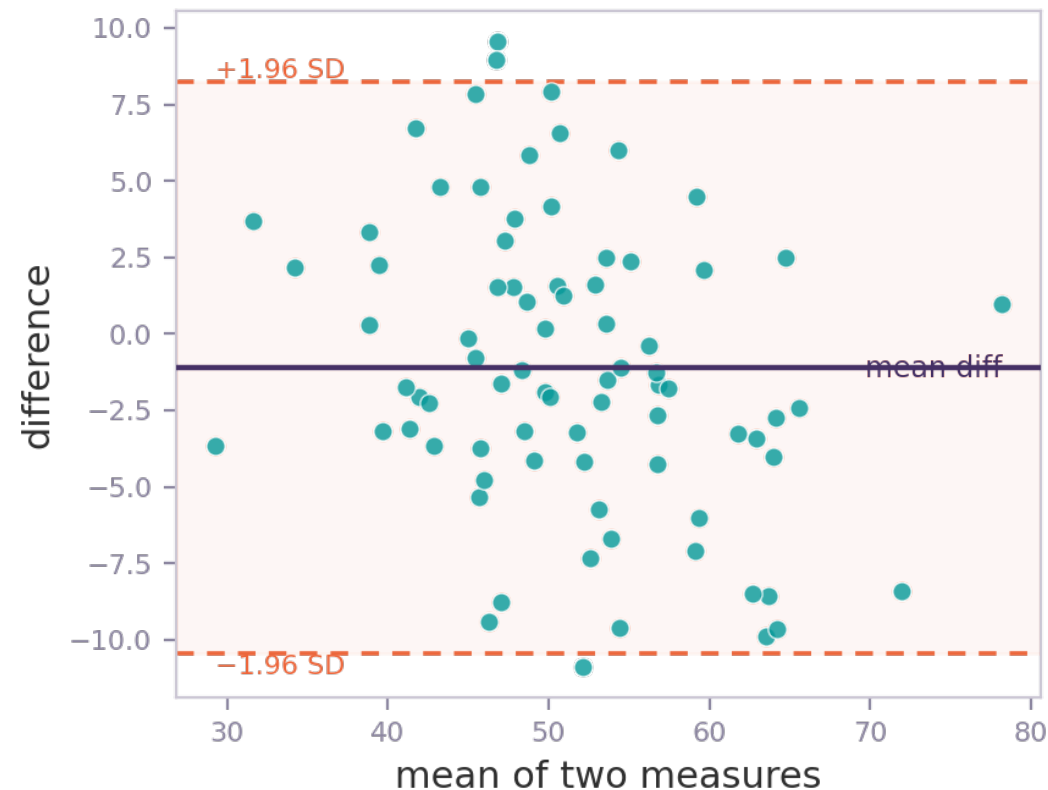
```
library(corr)

wide %>%
  select(TARGET, RELATEDA,
         RELATEDB, UNRELATED) %>%
  correlate() %>% shave() %>% fashion()
```



## Bland-Altman: agreement, not just association

Bland-Altman requires continuous data



### WHAT IT SHOWS

- Agreement between two methods or occasions: the difference plotted against the mean.
- Limits of agreement (mean difference  $\pm$  1.96 SD each with its own 95%CI) show how far paired measurements fall apart.
- A high correlation can still hide systematic bias but this plot can help reveal it. Pairs well with a good cup of tea and the ICC (Topic 2).

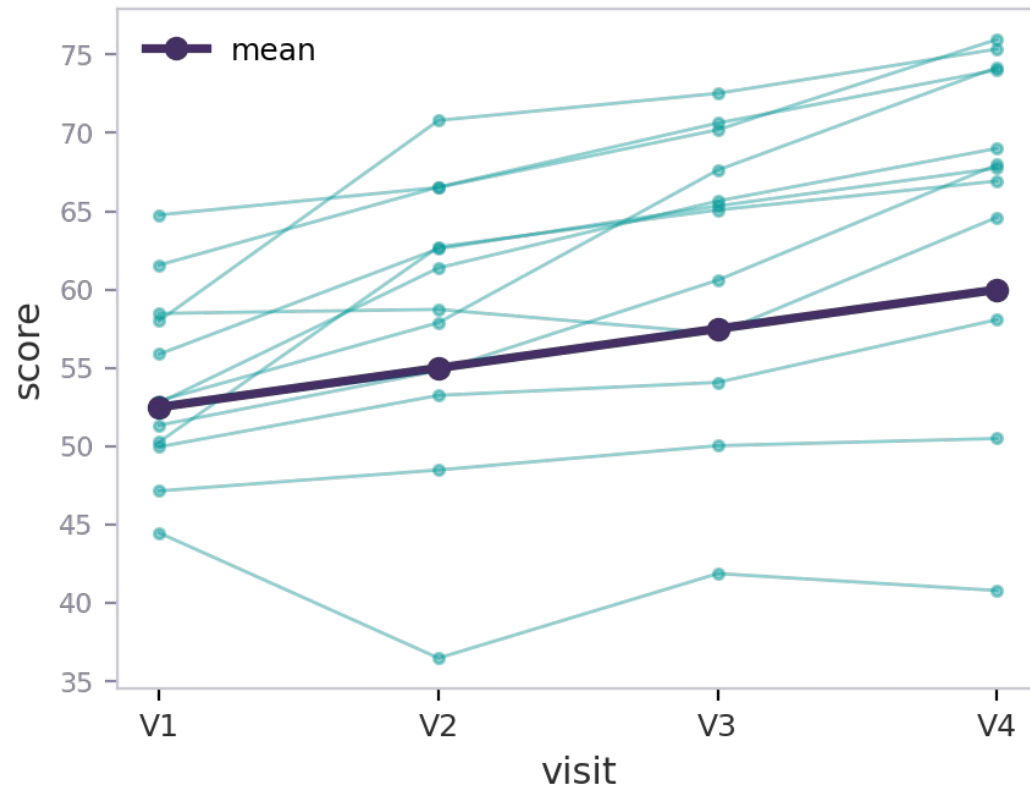
#### R • RUNS ON THE EXAMPLE DATASET

```
library(blandr)

# paired measurements per subject
blandr.draw(retest$Baseline, retest$Retest)
# bias 0.60    limits -13.9 to 15.1
```



## Spaghetti plot: individual change over time



### WHAT IT SHOWS

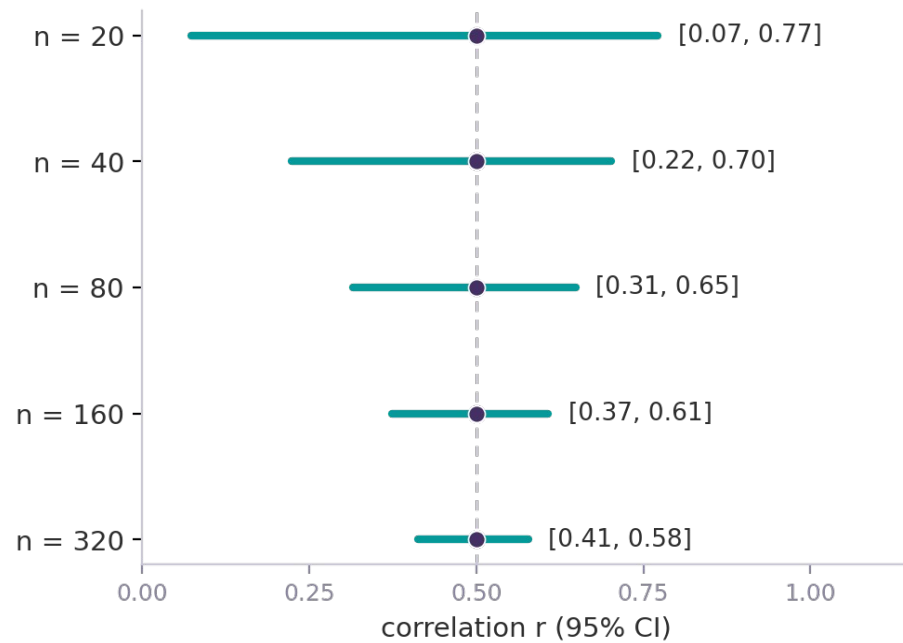
- Each subject's trajectory across visits, with the mean overlaid.
- Heterogeneity and individual change that a single correlation cannot show.
- A useful look before within / between modeling ([Topic 2](#)).

### R • RUNS ON THE EXAMPLE DATASET

```
# CDISC BDS is already long: plot directly
library(ggplot2)
adcorr %>%
  filter(PARAMCD == "SYMBUR") %>%
  ggplot(aes(AVISITN, AVAL, group = USUBJID)) +
  geom_line(alpha = 0.25) +
  stat_summary(aes(group = 1), fun = mean,
               geom = "line", linewidth = 1.4)
```



## Report the estimate with a confidence interval



*The 95% CI for  $r = 0.5$  narrows as the sample grows.*

- A confidence interval shows how precise the estimate is, which a lone  $r$  hides.
- Its width depends on the sample size, and on  $r$  itself (Fisher, 1921).
- Report  $r$  with its 95% CI in every correlation table.

### REPORTING TEMPLATE

**$r = 0.52$  (95% CI [0.41 ; 0.61]),  $n = 180$ .**



### IN PRACTICE

Non-overlapping 95% CIs indicate a real difference between two correlations. However overlapping CIs do not rule one out. A CI-for-the-difference is narrower than the eye suggests. If needed, two correlations can be formally compare to test the difference directly (Fisher  $z$  for independent samples; Steiger's test for dependent ones).



## Why we rarely use significance tests

A significance test for a correlation asks only whether  $r$  differs from exactly zero. In COA work that is almost never the question.

- With large samples, almost any correlation is “significant”, even a trivial one.
- With small samples, a clinically meaningful correlation can be “non-significant”.
- The useful questions (is it large enough, precise enough, as hypothesized?) are answered by the magnitude and the CI, not the p-value.



### WATCH OUT

A p-value rewards sample size, not just measurement quality (as it is partially a function of sample size). A large, precisely estimated correlation and a tiny one can carry the same p-value.



### IN PRACTICE

pre-specify the magnitude you need and report the estimate with its CI. If a p-value appears at all, keep it as secondary.



## Threshold benchmarks: a history, and a caution

There is no universal standard for a “weak” or “strong” correlation. Several conventions exist, and they disagree.

| Convention                           | Interpretation bands for $ r $                                                                      |
|--------------------------------------|-----------------------------------------------------------------------------------------------------|
| Cohen (1988)                         | 0.10 small · 0.30 medium · 0.50 large                                                               |
| Schober et al. (2018)                | <0.10 negligible · 0.10–0.39 weak · 0.40–0.69 moderate · 0.70–0.89 strong · $\geq 0.90$ very strong |
| Hinkle et al. (2003) / Mukaka (2012) | <0.30 negligible · 0.30–0.49 low · 0.50–0.69 moderate · 0.70–0.89 high · $\geq 0.90$ very high      |



### KEY INSIGHT

These began as general rules of thumb, not COA-specific laws. The same  $r$  can be “moderate” in one system and “weak” in another. Choose and justify one in advance, and interpret it against the measurement purpose, not as a verdict. Higher thresholds could be used where a method effect is thought to exist (two PROs completed in the same session) whereas more lenient values could be used when considering mixed modality correlations (PRO versus clinician administered performance-based outcome or a passively recorded digital measure). Typically, the more conservative estimates are expected by reviewers without a strong rationale



## Benchmarks for evaluating psychometric properties

| Property                  | What you correlate       | Common benchmark for $ r $ (full references in appendix)                                |
|---------------------------|--------------------------|-----------------------------------------------------------------------------------------|
| Internal consistency      | average inter-item $r$   | 0.15–0.50 (Clark & Watson, 1995); 0.20–0.40 (Piedmont, 2014)                            |
| Item quality              | corrected item–total $r$ | $\geq 0.30$ (Nunnally, 1994); $\geq 0.40$ (Fayers & Machin, 2015)                       |
| Test–retest / inter-rater | ICC                      | $\geq 0.70$ acceptable; 0.75–0.90 good (COSMIN; Koo & Li, 2016; Cicchetti, 1994)        |
| Convergent / divergent    | $r$ with a criterion     | $\geq 0.50$ convergent; $< 0.30$ divergent (COSMIN; Cappelleri, 2013)                   |
| Responsiveness            | change $r$               | $\geq 0.30$ (COSMIN; Revicki, 2008)                                                     |
| Anchor selection          | change $r$ with anchor   | $\geq 0.30$ (Revicki, 2008; Hays, 2005); $\geq 0.50$ ROC (Devji, 2020; Griffiths, 2023) |

What are COSMIN bands? Several thresholds above cite COSMIN (CONsensus-based Standards for the selection of health Measurement INstruments), an international initiative that defines quality criteria and rating thresholds for measurement properties, widely used as default benchmarks in PRO validation. The full reference for every citation above is given in the appendix. Where COSMIN guidance exists, this is generally the preferred threshold to adopt. However, in absence of this, the more conservative threshold is typically recommended without a strong rationale to do otherwise.



## Tabulating correlation results

Put the coefficient and its interval first; keep p-values secondary or omit them. Plots should support throughout.

| Comparison                                     | r    | 95% CI        | n   |
|------------------------------------------------|------|---------------|-----|
| Total score × global VAS (convergent)          | 0.62 | 0.54 to 0.69  | 180 |
| Domain A × Domain B (convergent)               | 0.45 | 0.34 to 0.55  | 180 |
| Total score × unrelated scale (divergent)      | 0.11 | −0.03 to 0.25 | 180 |
| Change in total × PGIC anchor (responsiveness) | 0.38 | 0.25 to 0.50  | 150 |

### FORMATTING

- Report r to two decimals with its 95% CI and n.
- Group convergent and divergent pairs together.
- State the hypothesized direction for each pair.
  - This can be added as a column to a results table
- Keep any p-values in a secondary column, or omit.



## Five things to carry forward



### Plot before you trust

The same  $r$  can come from very different data; the picture is the evidence. Scatter, matrix, Bland-Altman, and spaghetti each reveal something  $r$  hides.



### Report $r$ with a CI

A confidence interval shows the precision a single value hides.



### Significance is rarely the point

Size and precision answer the real questions, not the p-value.



### Benchmarks are conventions

No universal standard exists; pre-specify and justify the one you use.



### Match the benchmark to the property

Internal consistency, reliability, validity, responsiveness, and anchors each have their own.



## References

- Slide 58** Anscombe, F. J. (1973). Graphs in statistical analysis. *Am Stat*, 27(1), 17-21.
- Slide 61** Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 327, 307-310.
- Slide 63** Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1, 3-32.
- Slide 65** Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Erlbaum;
- Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesth Analg*, 126(5), 1763-1768;
- Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied Statistics for the Behavioral Sciences* (5th ed.);
- Houghton Mifflin. Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*, 24(3), 69-71.
- Slide 66** Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309-319;
- Piedmont, R.L. (2014). Inter-item Correlations. In: Michalos, A.C. (eds) *Encyclopedia of Quality of Life and Well-Being Research*. Springer, Dordrecht;
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill;
- Fayers, P. M., & Machin, D. (2015). *Quality of life: the assessment, analysis and reporting of patient-reported outcomes*; John Wiley & Sons;
- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine*, 15(2), 155-163;
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284-290;
- Prinsen, C. A. C., et al. (2018). COSMIN guideline for systematic reviews of PROMs. *Qual Life Res*, 27, 1147-1157;
- Cappelleri, J. C., et al. (2013). *Patient-Reported Outcomes: Measurement, Implementation and Interpretation*. CRC Press;
- Revicki, D., et al. (2008). Determining responsiveness and minimally important differences for PROs. *J Clin Epidemiol*, 61(2), 102-109;
- Hays, R. D., Farivar, S. S., & Liu, H. (2005). Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD*, 2(1), 63-67;
- Coon, C. D., & Cappelleri, J. C. (2016). Interpreting change in scores on PRO instruments. *Ther Innov Regul Sci*, 50(1), 22-29;
- Devji, T., et al. (2020). Credibility of anchor-based estimates of minimal important differences. *BMJ*, 369, m1714;
- Griffiths, P., Sims, et al., (2023). How strong should my anchor be for estimating group and individual level meaningful change? A simulation study assessing anchor correlation strength and the impact of sample size, distribution of change scores and methodology on establishing a true meaningful change threshold. *Quality of Life Research*, 32(5), 1255-1264. (simulation).



TOPIC 5

# Challenges and solutions

What distorts a correlation, and what to do about it



## What can distort a correlation

Even the right coefficient on the right question can mislead if the data work against it. The common threats, and where they bite:



### Too few subjects

*Unstable estimate, wide CIs*



### Missing data

*Biased or inconsistent estimate*



### Restricted range, floor / ceiling

*Attenuated correlation*



### Measurement error

*Attenuated correlation*



### Recall bias

*Distorted anchor or rating*



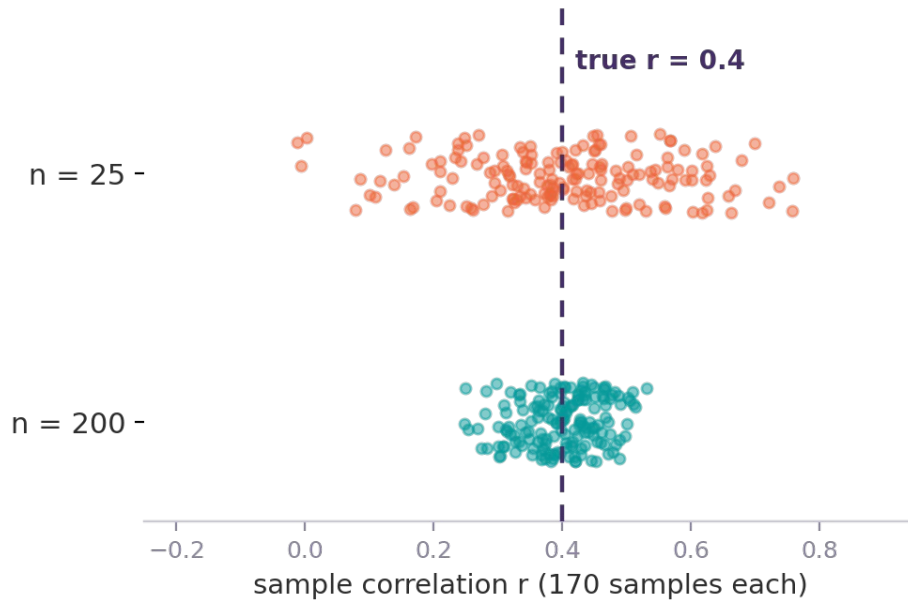
### Change vs cross-sectional

*Not directly comparable*

***Most fixes are about design and reporting, not a cleverer coefficient.***



## Too few subjects, too much noise



With  $n = 25$ , the estimate swings widely around the truth; with  $n = 200$  it settles.

- A small sample gives an unstable  $r$  that can be far from the true value.
- Confidence intervals are wide, so small studies rarely distinguish one correlation from another.
- There is no universal minimum, but precision improves steadily with  $n$  (Schönbrodt & Perugini, 2013).



### WHAT TO DO

Plan for the precision you need around the expected  $r$ , not just for a statistically significant result. Report confidence intervals and treat correlations from small samples as provisional.



## Missing data bends the estimate

How missingness is handled can change a correlation as much as the data themselves. The two default deletion rules trade off differently.

### LISTWISE DELETION

Drops every subject with any gap. One consistent matrix – every  $r$  on the same cases – but it sacrifices power (smaller  $n$ , wider CIs) and biases  $r$  unless data are MCAR.

### PAIRWISE DELETION

Uses every available pair for each correlation. Retains more data and power, which helps with a matrix of items, but each cell rests on a different subset and  $n$ , so the matrix can be inconsistent – even non-positive-definite, which breaks factor analysis.

### MISSINGNESS MECHANISM

**MCAR** (Missing Completely At Random): unrelated to any value.

**MAR** (Missing At Random): explained by observed data.

**MNAR** (Missing Not At Random): related to the missing value itself.

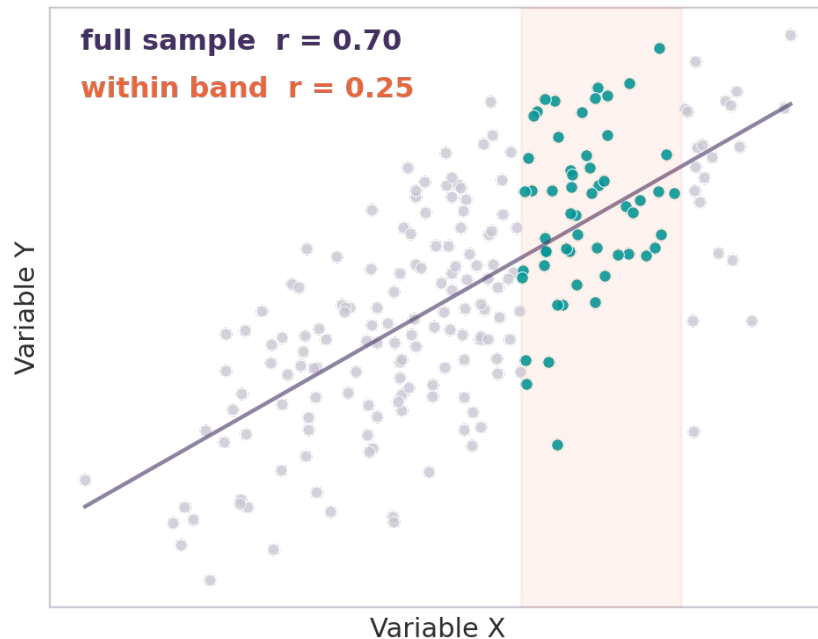


### WHAT TO DO

Report the amount and pattern of missingness. In measurement-property work the usual preference is not to impute: imputing here can bias the property estimate, and that bias compounds when the same scores feed efficacy analyses. Prefer observed data with a documented, pre-specified deletion rule, and never compare correlations computed on different subsets.



## When variability disappears, so does $r$



*The same data: a strong full-sample correlation nearly vanishes inside a narrow band.*

*The same shrinkage happens if Y's range is restricted instead of X's.*

- A narrow spread on either variable attenuates  $r$ , even when the underlying relationship is strong (Sackett & Yang, 2000).
- Common in practice: when inclusion criteria enrol a narrow band – for example only severe patients – the baseline range is restricted by design.
- Floor and ceiling effects, and homogeneous samples, compress the usable range the same way.

### WHY IT HAPPENS

$r$  is the covariance of two variables standardized by their standard deviations. If a variable barely varies, there is little covariance to detect, so  $r$  shrinks toward zero whatever the true relationship.



### WHAT TO DO

Inspect the spread and distributions before correlating. Do not compare  $r$  across samples with different ranges, and for change, make sure enough patients actually change.



## Measurement error pulls correlations toward zero

Unreliable measures attenuate the observed correlation below the true relationship between the constructs. The noisier the measures, the more  $r$  is pulled toward zero.

### Correction for attenuation

$$r_{\text{true}} \approx r_{\text{obs}} / \sqrt{r_{XX} \cdot r_{YY}} \quad \text{e.g. } 0.50 / \sqrt{(0.80 \cdot 0.80)} = 0.63$$

where  $r_{XX}$  and  $r_{YY}$  are the test-retest reliabilities. Use an ICC and state its form: a Pearson correlation between occasions ignores systematic shift and under-corrects.

*Spearman (1904); Nunnally & Bernstein (1994).*



### WATCH OUT

The disattenuated value estimates the latent correlation, but it is sensitive to the reliability estimates and can over-correct, sometimes exceeding 1. Treat it as an upper bound, not a result to report on its own.



### WHAT TO DO

Report the observed correlation and the reliabilities used. If you disattenuate, show the corrected value alongside the observed one. While this formula is useful, proper measure development ensures more reliable measures - which is preferable to estimating corrected value.



## Anchors: external references for interpreting change

Anchors recur throughout responsiveness and meaningful-change work, so they are worth defining. An anchor is a separate, interpretable measure used to give meaning to scores or change on the target measure.

### WHAT IT IS

Typically a global rating – the Patient Global Impression of Change (PGIC) or Severity (PGIS) – or a clinical milestone that patients and clinicians readily understand.

### WHAT IT IS FOR

Anchors underpin responsiveness evidence and the estimation of meaningful change. The change correlation between target and anchor is the central piece of evidence.

### WHAT MAKES A GOOD ANCHOR

- Appreciably correlated with change on the target. The cited  $\geq |0.30|$  is a low floor, not a target: higher is preferred.
- Itself interpretable and meaningful to patients or clinicians.
- A defensible recall period for the concept being rated.
- Measure the same concept as the measure of interest.



### WATCH OUT

A weak or biased anchor undermines every conclusion drawn from it – see recall bias next.

The anchor's validity rests on qualitative and quantitative evidence, not psychometrics alone, even if it's treated as the quantitative gold standard.



## Recall bias in retrospective ratings

Anchors and some PROs ask patients to recall a past state or summarize a period. Recall is reconstructed, not replayed.

- Retrospective ratings are colored by the patient's current state and by recent or peak experiences (Stull et al., 2009).
- Longer recall windows make the bias worse.
- A biased anchor distorts the very change correlation it is meant to support.

**Examples:** PGI-C and CGI-C



### WHAT TO DO

Keep recall windows short and prefer momentary or prospective assessment where the design allows it. Choose anchors with a defensible recall period for the concept.

**Worth remembering:** a global rating of change asks patients to compare now with a past they may no longer represent accurately. Further, recall period can change as time goes on. Interpret anchor-based change correlations with that limit in view.



## Change and cross-sectional correlations are not comparable

### CROSS-SECTIONAL

- Cannot demonstrate responsiveness; a strong  $r$  may reflect a shared concept, not sensitivity to change.
- Response shift: patients' internal standards or how they understand the concept can change over time, so cross-sectional validity need not imply responsiveness.
- Driven by differences between patients, not change within them.



### WHAT TO DO

Do not expect change and cross-sectional correlations to match. Use cross-sectional for construct validity and change for responsiveness, pre-specify each, and be wary of baseline-change correlations.

### CHANGE SCORE CORRELATION

- Usually lower, because error compounds across two measurements.
- Limited improvement or worsening reduces the change variance to correlate.
- Strongly dependent on the anchor, and on floor or ceiling effects and reliability of the change score.



### WATCH OUT

Regression to the mean: extreme baseline scores drift toward the average, so change correlates negatively with baseline before any treatment effect. Baseline sits on both sides of the subtraction, so this is close to automatic (Clifton & Clifton, 2019).

A study that measures different samples of patients at multiple time points (repeated cross-sections) is not the same as a longitudinal study that tracks the same patients over time—even though both “have multiple time points.”



## A checklist for a trustworthy correlation



### Plan for precision

Size the study around the expected  $r$ , and report confidence intervals.



### Handle missing data well

Report its rate and pattern; in the context of psychometric work prefer observed data with a pre-specified rule over imputing.



### Check the spread

Watch for restricted range, floor, and ceiling before interpreting  $r$ .



### Respect measurement error

It attenuates  $r$ ; correct for reliability only with caution, and show both values.



### Mind anchors and recall

Define the anchor, keep recall windows short, and pick a defensible period.



### Keep the levels apart

Do not compare change with cross-sectional  $r$ , and beware [regression to the mean](#).



## References

- Slide 68** Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *J Res Pers*, 47(5), 609-612.
- Slide 73** Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.
- Slide 70** Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *J Appl Psychol*, 85(1), 112-118.
- Slide 71** Spearman, C. (1904). The proof and measurement of association between two things. *Am J Psychol*, 15, 72-101;  
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill (correction for attenuation).
- Slide 76** Revicki, D., et al. (2008). Determining responsiveness and minimally important differences for PROs. *J Clin Epidemiol*, 61(2), 102-109;  
Coon, C. D., & Cappelleri, J. C. (2016). Interpreting change in scores on PRO instruments. *Ther Innov Regul Sci*, 50(1), 22-29;  
Devji, T., et al. (2020). Credibility of anchor-based estimates of minimal important differences. *BMJ*, 369, m1714.
- Slide 73** Stull, D. E., Leidy, N. K., Parasuraman, B., & Chassany, O. (2009). Optimal recall periods for patient-reported outcomes: challenges and potential solutions. *Curr Med Res Opin*, 25(4), 929-942.
- Slide 76** Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *Int J Epidemiol*, 34(1), 215-220;  
Clifton, L., & Clifton, D. A. (2019). The correlation between baseline score and post-intervention score, and its implications for statistical analysis. *Trials*, 20, 43.  
Sprangers, M. A. G., & Schwartz, C. E. (1999) Integrating response shift into health-related quality of life research. *Soc Sci Med*, 48(11), 1507-1515.



Correlations

## TOPIC 6

# Further Relevant Concepts

Concepts that are useful for alongside correlation and change how a coefficient should be interpreted.

# What anchors are available and should they be used

**DEFINITION:** External measure, distinct from the PRO being evaluated, used to give meaning to a score or a change in score on that PRO

## WHEN TO USE

- To estimate a minimal clinically important change from the patient's perspective (Minimal Within-Patient Change).
- To provide evidence of the "responsiveness" (sensitivity to change) of a PRO/COA.

## METHODS

Calculate the correlation between the change in the target instrument and the change (or level) of the anchor.

Verify that the change-anchor correlation is "appreciable": commonly threshold = 0.30.



## WATCH OUT

- The anchor must itself be interpretable and understandable to patients or clinicians.
- Select an anchor theoretically linked to the measured concept, pre-specified prior to analysis.
- A weak or biased anchor invalidates any conclusion that depends on it.

## List of main anchors:

- State anchors: CGIS, PGIS,
- Anchors of change: CGIC, PGIC



## What are Latent variables

### DEFINITION:

A variable that cannot be measured directly but is inferred from other observed variables.

**USED** when a COA instrument is built on the assumption that a score reflects an underlying theoretical construct

### How to construct and Verify:

- EFA, a statistical method that explores, without strong a priori hypotheses, how many latent variables underlie a set of items and which items are associated with which factor.
- CFA is a method that tests a pre-specified structure of latent variables and their hypothesized item loadings against observed data by evaluating model fit.

### EXAMPLE:

- The total score on a quality-of-life questionnaire is an observed variable; the *quality of life* it is intended to measure is the latent variable.
- A 10-item depression scale observes 10 manifest variables linked to a single latent variable: *depression severity*.



### WATCH OUT

- The latent variable is never measured directly, only estimated.
- A latent variable is a modeling assumption, not an empirical fact
- The latent variable remains a construct, not an object measurable in itself.



## INTERPRETATION BONUS

# $R^2$

### Definition:

$R^2$  is the square of the Pearson correlation coefficient  $r$ . It represents the proportion of the variance of one variable explained by the other.

### Uses:

- To correctly compare two correlations of different magnitudes.
- As an alternative effect size to communicate the importance of an association. Often more intuitive than  $r$  for a non-statistical audience.
- In multiple regression, to assess the total variance explained by a set of predictors.
- % of shared variance.



### WATCH OUT

- $R^2$  loses the sign of the association. Always present  $r$  alongside  $R^2$ .
- $R^2$  is not linearly comparable to  $r$ .
- Do not interpret  $R^2$  as causality

### EXAMPLE:

- Result of correlation between PRO and an anchor is  $r = 0.7$ .  $R^2 = 0.49$ , so 49% shared variance between the PRO and the anchor.
- A first result of correlation obtained is  $r = 0.5$ ; a second one is  $r = 0.25$ . The  $R^2$  are 0.25 and 0.0625. The first association is 4 times stronger in terms of explained variance, not twice as strong.



# Scoring direction

### Scoring direction of the variable:

- describes how a score is coded in terms of clinical meaning: does a higher score indicate a better or worse health status?
- This choice of convention (e.g., 0 = no symptoms vs. 0 = best health) directly affects the sign of the observed correlations.

### Context of importance:

- Whenever the sign of a correlation between two COAs, or between a COA and an anchor is interpreted.
- Before comparing correlations across multiple instruments (scoring directions must be harmonized; otherwise, the signs become inconsistent and misleading).

### Solutions:

- Explicitly document the scoring direction for each instrument.
- Before calculating or interpreting a correlation between two measures, verify that their scoring directions are known and consistent (if necessary, reverse one of the scores, e.g., by multiplying by  $-1$  or using reverse-coding).
- In a validation report, always state the expected direction of the correlation based on the clinical meaning of each scale, rather than just the raw statistical sign.



### WATCH OUT

- Note if a score has been reversed
- Always specify the scoring direction in the legend of any correlation table or matrix.
- Some instruments change conventions between versions check the scoring direction before comparing studies over time.



# Simpson's Paradox and Lord's Paradox

## DEFINITION:

A pattern seen within every subgroup can reverse, weaken, or disappear once the subgroups are combined (Simpson, 1951). In repeated-measures or change-score analyses, adjusting for baseline can flip the direction of an effect (Lord, 1967) — this specific case is often called Lord's Paradox.

## WHEN

when a variable (e.g. disease severity, treatment group, baseline score) differs across subgroups and is correlated with both the variables being compared.

## How to check:

- Compute the correlation within each relevant subgroup, not only in the pooled sample.
- If subgroup and pooled results disagree, report both and investigate the subgrouping or baseline variable before interpreting the pooled figure.
- For change scores, check whether the correlation with baseline changes sign once the analysis is restricted to a homogeneous subgroup (Simpson, 1951; Lord, 1967).

## EXAMPLE:

- A treatment appears to reduce symptom severity in the pooled sample, but within each disease-severity subgroup it shows no effect. The pooled correlation is driven by which subgroup patients fall into, not by the treatment itself.
- Two clinicians' baseline severity ratings correlate positively across patients, but after adjusting for each patient's own baseline, the correlation between their change scores reverses sign.



## WATCH OUT

- A pooled correlation or mean difference can disagree with the result seen in every subgroup taken separately.
- Checking only the aggregate can hide, or manufacture, a real reversal.
- Adjusting for baseline can change the sign of a treatment or change effect, depending on how the compared groups differ at baseline (Lord's Paradox).



# Regression to the mean

## DEFINITION:

Patients selected for having an extreme score (very high or very low) tend to score closer to the average on a later measurement, even with no real change and no intervention (Barnett, van der Pols, & Dobson, 2005).

## WHEN

when patients are selected or enrolled based on an extreme baseline score, and again whenever a baseline-to-change correlation is interpreted.

## How to guard against it:

- Use a control or comparison group selected the same way, so both groups regress equally.
- Do not rely on a single extreme baseline measurement.
- Be cautious interpreting any correlation between a baseline score and its own change score.

## EXAMPLE:

A trial enrolls only patients with the most severe baseline symptom scores. At follow-up, average scores improve, but part of that improvement would have happened without treatment, simply because extreme baseline scorers regress toward the mean.



## WATCH OUT

- Regression to the mean can manufacture a negative baseline-to-change correlation even with no true change: extreme scorers drift toward the average.
- It is not measurement error alone, it also comes from imperfect reliability and from within-person fluctuation over time.
- It can be mistaken for a genuine treatment effect if patients are selected on an extreme baseline score and there is no control group.



## Corrected item-total correlation

### DEFINITION:

The correlation between one item's score and the sum of the other items in the scale, excluding that item itself. "Corrected" refers to that exclusion, without it, the item would be partly correlated with itself.

### USED

during scale development or refinement, to check whether each item contributes to the same construct as the rest of the scale.

### How it differs from the uncorrected version:

- Uncorrected item-total correlation includes the item itself in the total, which inflates the estimate, especially in short scales, where one item is a larger share of the total.
- The corrected version is the one commonly reported and the one typically meant by "item-total correlation".
- Related to, but distinct from, inter-item correlations: item-total looks at one item against the rest combined, inter-item looks at item pairs.

### EXAMPLE:

- A 10-item scale: item 7's corrected item-total correlation is  $r = 0.15$ , well below the other items ( $r = 0.40-0.60$ ).
- This flags item 7 for review, it may be poorly worded, measure a different construct, or need rescoring. This result is not an automatic instruction to delete it.



### WATCH OUT

- A low value can mean a bad item, but can also mean the scale is multidimensional and this item belongs to a different factor
- A high value can hide item redundancy: two near-identical items will both show strong corrected item-total correlations without adding information.
- The value is sample-dependent and does not account for item difficulty or location, so it should not be the only criterion for retaining or removing an item.



## Reliability and attenuation

### DEFINITION:

The reliability of a measure is the square of the correlation between its observed score and the true score.

Every observed score is split into two parts:  $x = \tau + \epsilon$ , where  $x$  is the observed score,  $\tau$  (tau) is the true, latent score, and  $\epsilon$  (epsilon) is measurement error.

Reliability is defined from this model as  $r(x, x') = r(x, \tau)^2$

The correction-for-attenuation formula estimates the true-score correlation:  $r(\tau_x, \tau_y) = r(x, y) / \sqrt{(r(x, x') \times r(y, y'))}$  (Spearman, 1904).

### WHY IT MATTERS

Every correlation computed on observed scores ( $x, y$ ) is a correlation between two noisy proxies for the constructs you actually care about ( $\tau_x, \tau_y$ ). Error in either measure pulls the observed correlation toward zero relative to the true relationship.

### EXAMPLE:

Two PROMs correlate at  $r = 0.50$ . If each has a reliability of  $r(x, x') = 0.70$ , the attenuation-corrected estimate of the true-score correlation is  $0.50 / \sqrt{(0.70 \times 0.70)} \approx 0.71$ .



### WATCH OUT

- A correlation that looks weak may reflect low reliability, not a weak true relationship — check reliability before concluding the constructs are unrelated.
- Lord & Novick (1968, Ch. 3) show this only holds if  $r(x, y) \leq \sqrt{r(x, x')}$ ; a corrected value above 1.0 signals a violation of the model, not a stronger relationship.
- When the observed correlation is pushed toward 1 by learning (PerfO), this is the opposite problem from attenuation, and the correction-for-attenuation formula does not apply, since the extra correlation isn't due to reliability at all.



## References

- Slide 83** Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110(2), 203-219.
- Slide 84** Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin*, 97(2), 307-315.
- Slide 85** Streiner, D. L., Norman, G. R., & Cairney, J. (2024). *Health Measurement Scales: A practical guide to their development and use* (6th ed.). Oxford University Press.
- Slide 86** Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society B*, 13(2), 238-241; Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68(5), 304-305.
- Slide 87** Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, 34(1), 215-220.
- Slide 88** Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill.
- Slide 89** Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Menlo Park.