gesis Leibniz Institute for the Social Sciences

Harmonizing survey data across different survey modes

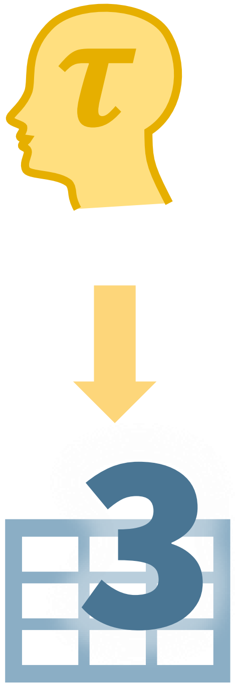
Dr. Ranjit K. Singh (ranjit.singh@gesis.org)

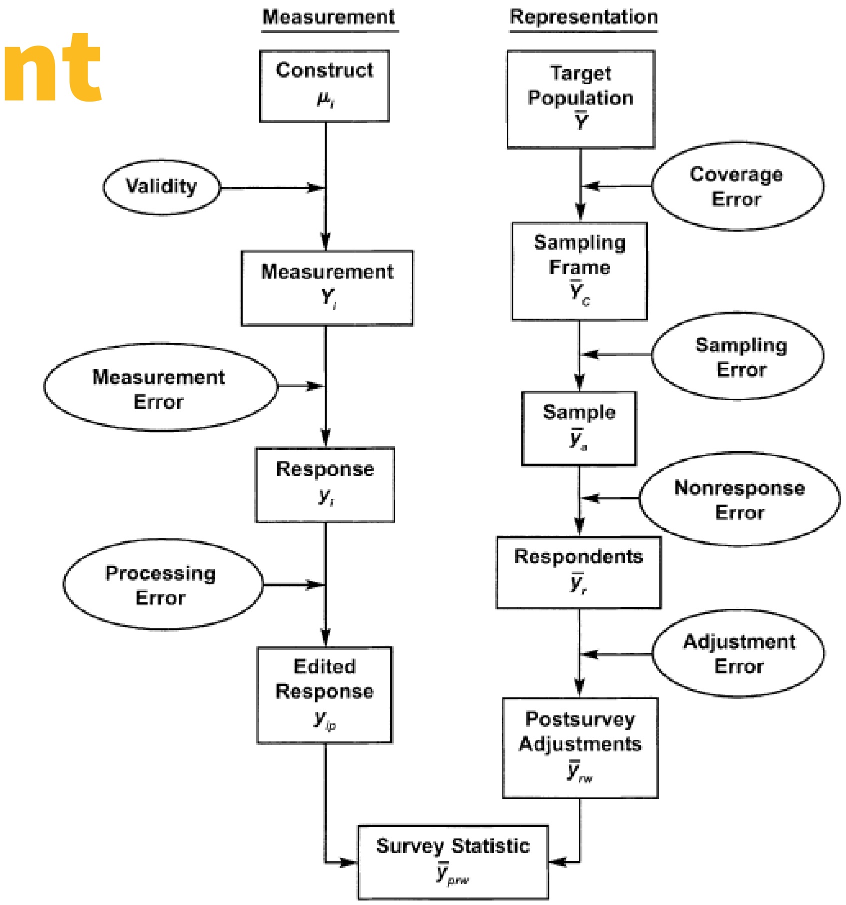
ESS & NatCen Survey Methodology Seminar ***2022-10-19***

Leibniz Association

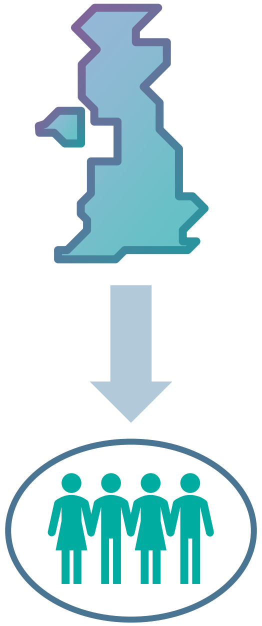
Total Survey Error: *The **smaller**, the better!*

Measurement



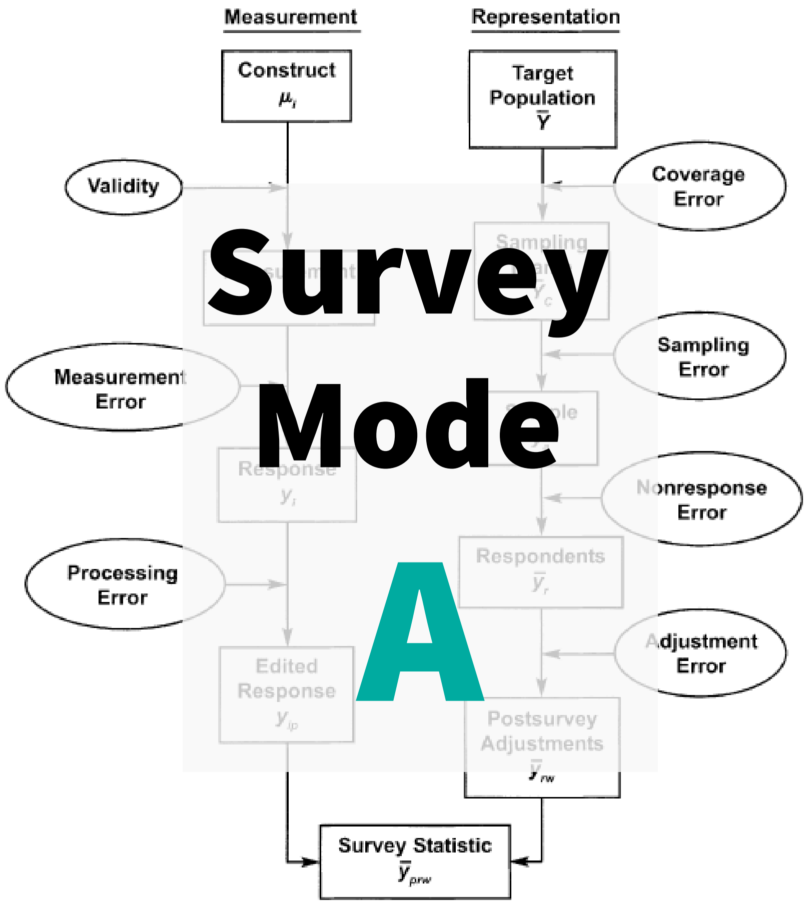


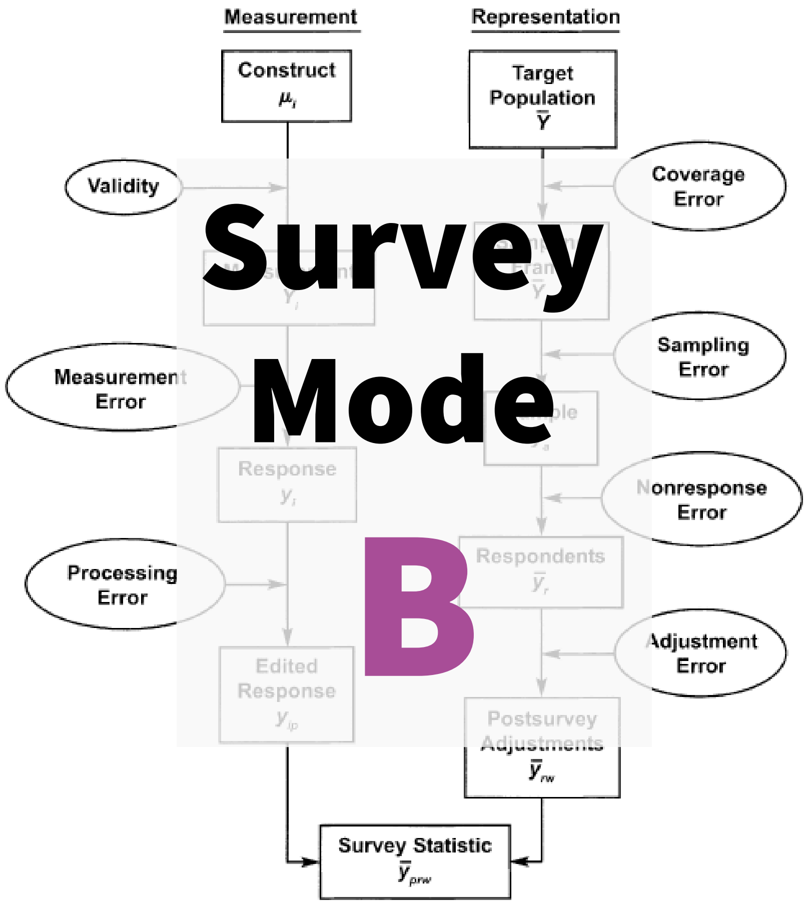
Representation



Groves et al., 2009

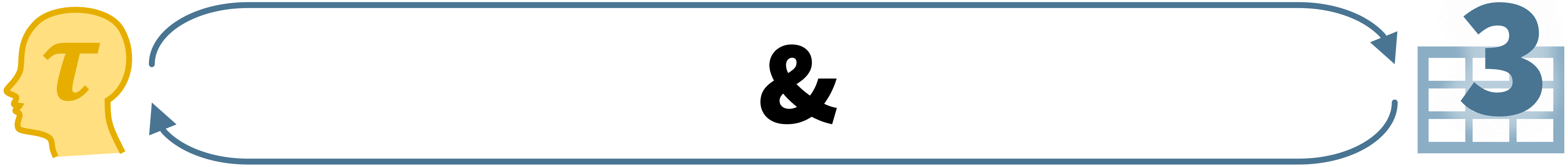
Survey Error(s): *The more* ***similar****, the better!*





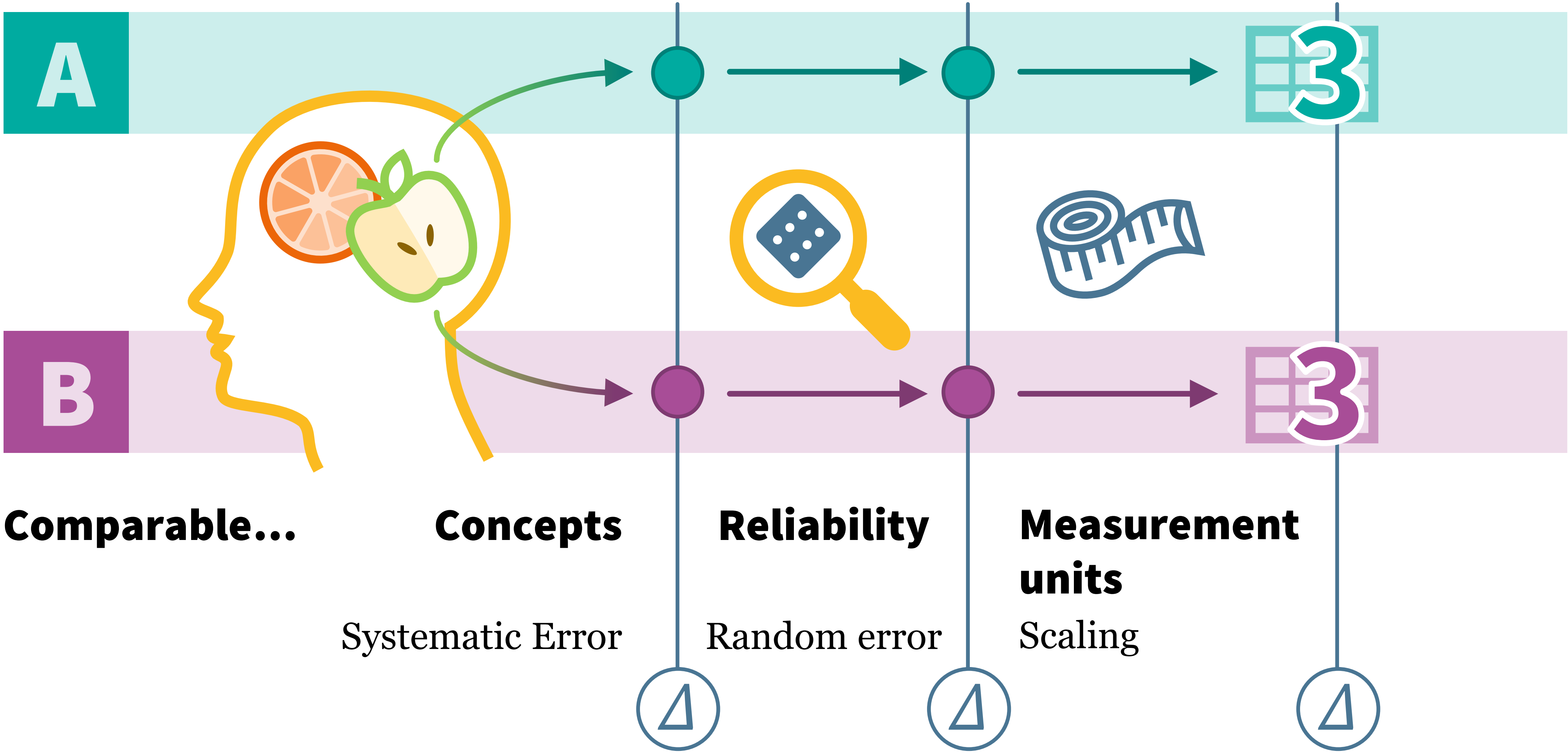
Comparable Measurement

The Respondents with the **same true score**
should give the same response (on average),
regardless of the survey mode.



The **same response score** in our data
should allow the same inferences about the respondent,
regardless of the survey mode.

Components of Comparability



Comparable...

Concepts

Systematic Error

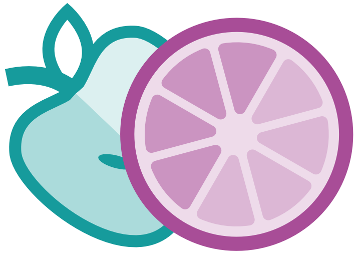
Reliability

Random error

Measurement units

Scaling

Comparable Concepts



The first and most fundamental issue in comparability:
Do we measure the same concept?

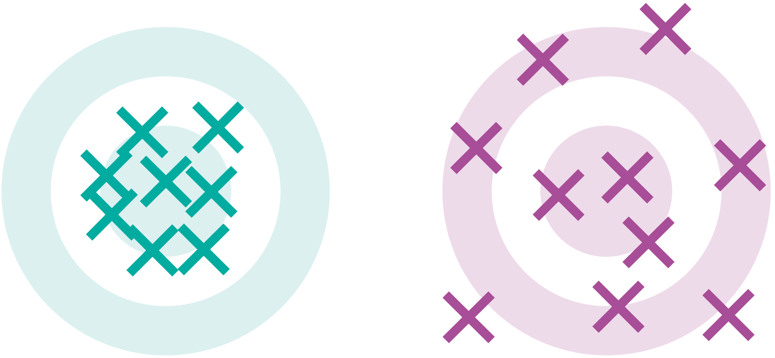
- Between different survey modes, **substantive differences in question understanding** are unlikely.
- However, survey modes may **contaminate measurement** with mode specific **systematic errors**

Examples:

- More **socially desirable responding** when an Interviewer is present?
- Greater **respondent burden** in one mode may intefere with memory retrieval

Comparable reliabilities

- Random error is **non-systematic error variance**
  Reliability is the other side of the same coin
- **Attenuation**
  The **less reliable** our measurement,
  the **lower are correlations** in our analyses
- If **survey modes** lead to **different reliabilities**, substantive correlations are spuriously lower in one mode than the other

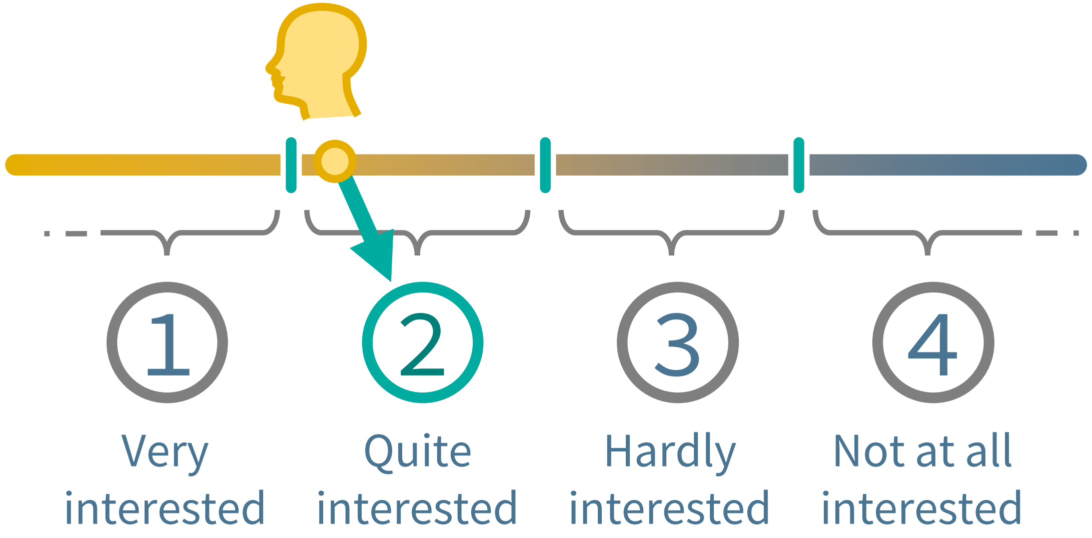


Example:

A survey switches its mode. The new mode leads to higher random error and thus lower reliability. Now we find that political interest suddenly predicts political participation less after the mode switch. A methodological artifact due to attenuation!

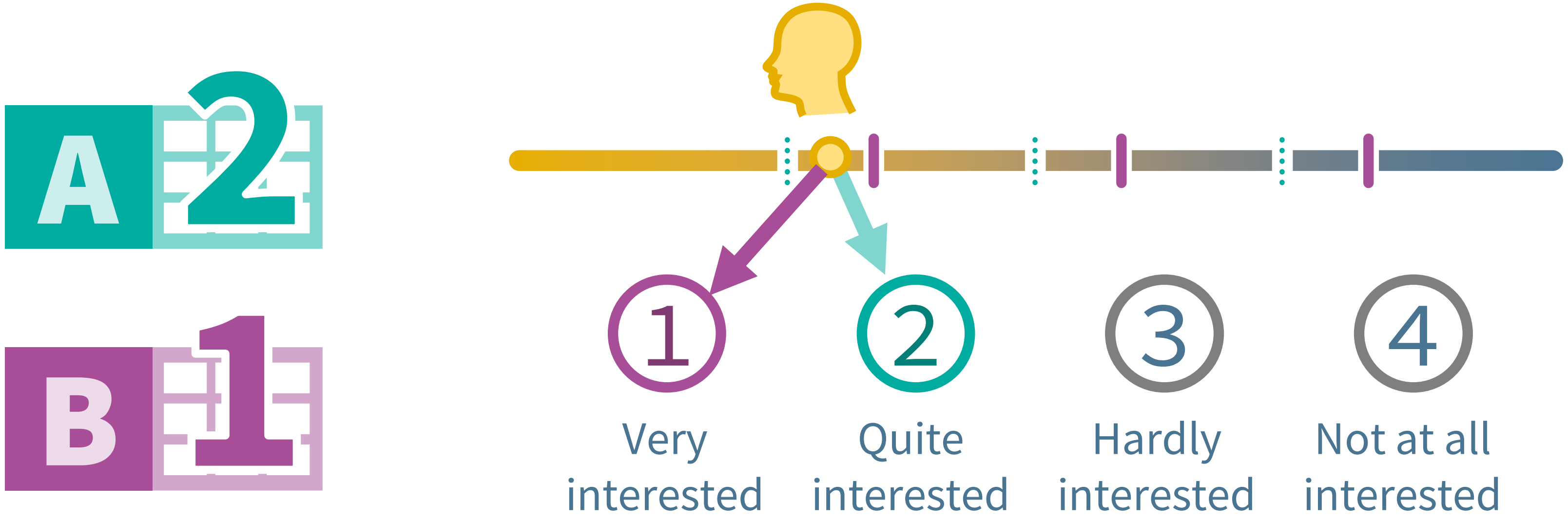
Comparable Measurement Units

- Many survey questions capture a **continuous concept** in an **ordinal (or pseudo-metric) measurement scheme**

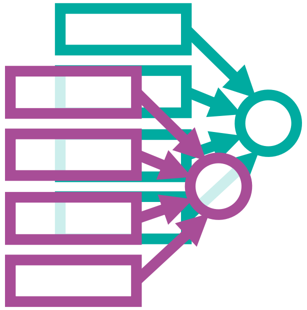


Comparable Measurement Units

- Many survey questions capture a **continuous concept** in an **ordinal (or pseudo-metric) measurement scheme**
- This **mapping may change between different survey modes**

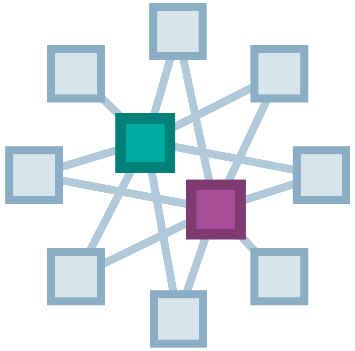


Four Ideas to assess (and mitigate) mode comparability issues



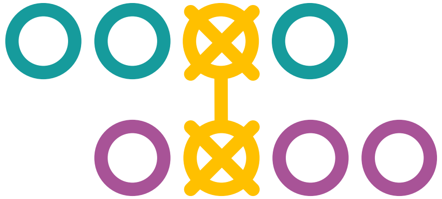
Formal Measurement Invariance

MGCFA



Concepts and Reliability

R-Alerting and comparative attenuation

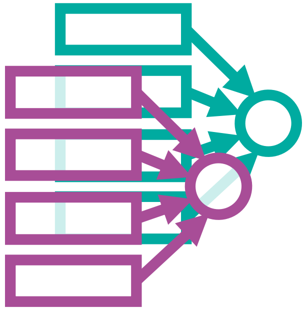


Aligning measurement units

OSE-RG

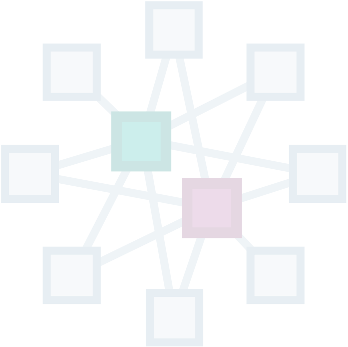
Generalizable Mode Effects

MTMM Meta-Analysis with SQP



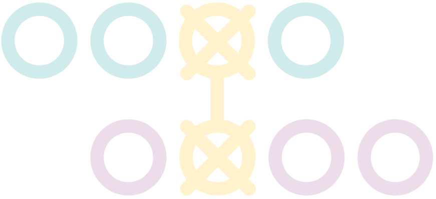
Formal Measurement Invariance

MGCFA



Concepts and Reliability

R-Alerting and comparative attenuation



Aligning measurement units

OSE-RG

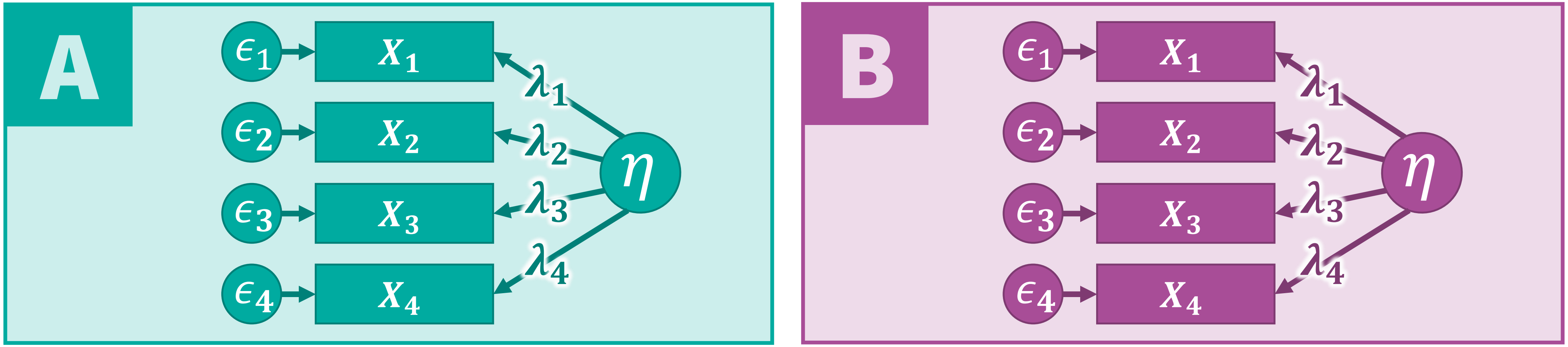
Generalizable Mode Effects

MTMM Meta-Analysis with SQP

gesis Leibniz Institute for the Social Sciences

MGCFA to assess Measurement Invariance (MI)

- CFAs assess construct structure, reliability, and measurement units
- MGCFAs then do the same for modes **A** and **B**, and then compare if the measurement instrument behaves differently



MGCFA to assess Measurement Invariance (MI)

- ✓ MGCFAs are a **formal and powerful framework** for comparability
- ✓ With one approach, we can cover **several comparability components at once**

- × Only applicable to **psychometric Multi-Item Instruments**
- × **Interpreting (MG-)CFA results** can be complex
- × They are **not a panacea**. E.g., MGCFAs can be blind to some errors that affect all items equally.

MGCFA to assess Measurement Invariance (MI)

Examples:

Davidov, E., Depner, F. Testing for measurement equivalence of human values across online and paper-and-pencil surveys. *Qual Quant* **45**, 375–390 (2011). https://doi.org/10.1007/s11135-009-9297-9

Roberts, C., Sarrasin, O., & Ernst Stähli, M. (2020). Investigating the Relative Impact of Different Sources of Measurement Non-Equivalence in Comparative Surveys. *Survey Research Methods*, 14(4), 399-415. https://doi.org/10.18148/srm/2020.v14i4.7416

Formal Measurement Invariance

MGCFA

Concepts and Reliability

R-Alerting and comparative attenuation

Aligning measurement units

OSE-RG

Generalizable Mode Effects

MTMM Meta-Analysis with SQP

Construct / Criterion Validation

- **Measurement instruments** are usually **validated** by correlating them to **related (or intentionally unrelated) concepts**
- Here, we do the same for **two modes**
  (Ideally in a random mode experiment)

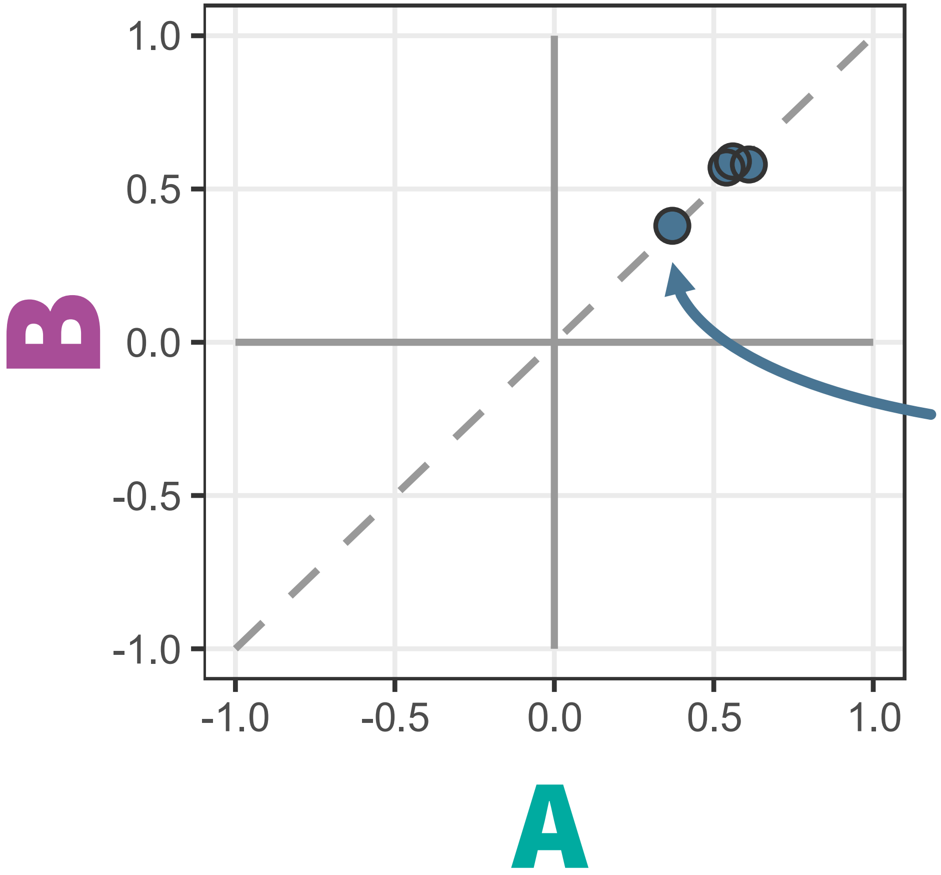
Construct Validity		**Criterion Validity**
Convergent	Divergent	Concurrent (& Predictive)
High correlations with **related concepts**	Low correlations with **unrelated concepts**	High correlations with **relevant outcomes**

Example: Interest in Politics

	A	B
Political Interest correlated with:	r_A	r_B
Interest in TV news	.37	.38
Interest in political TV shows	.61	.58
Understanding of the important political issues facing Germany	.54	.57
How often do you discuss politics?	.56	.59

If **modes** **A** and **B** **work similarly**, we would expect **similar correlations** in both modes (row-wise)

Summarising Validity Correlations



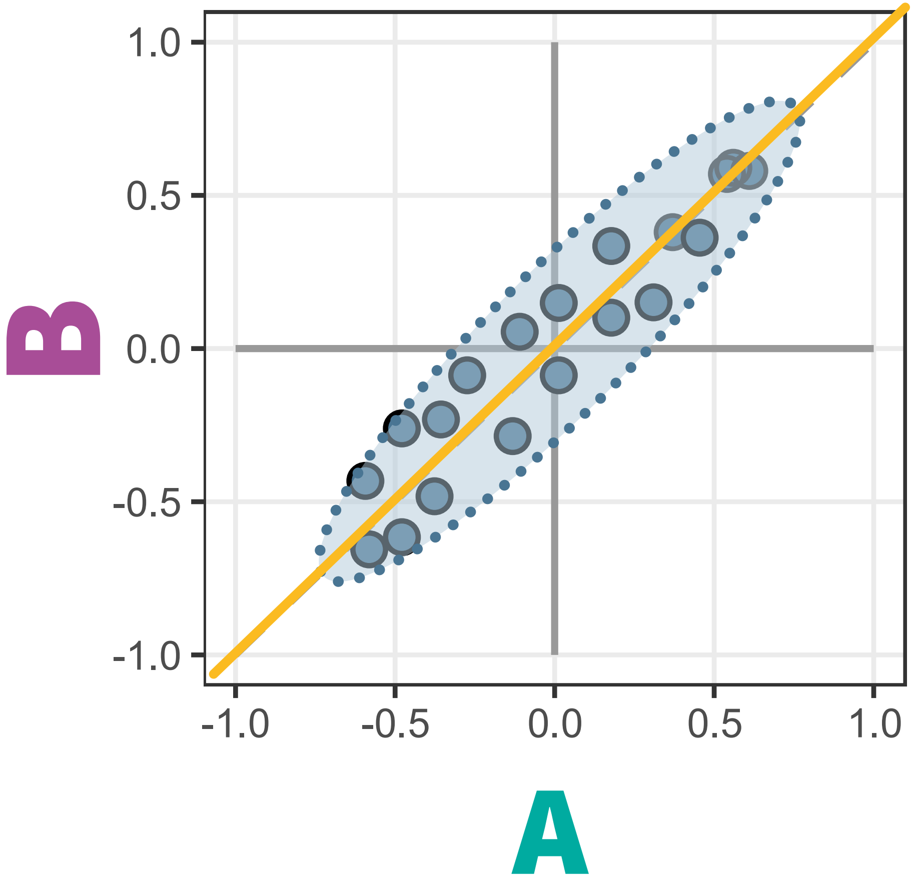
Correlation of Correlations

$r_{Alerting-CV} = .96$

e.g., „Interest in TV news“

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. Journal of Personality and Social Psychology, 84(3), 608–618. https://doi.org/10.1037/0022-3514.84.3.608

Summarising Validity Correlations

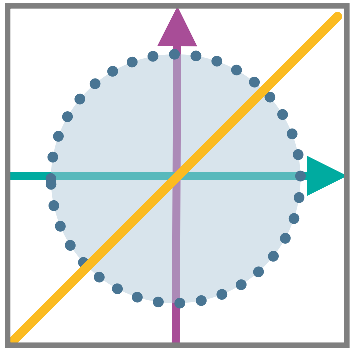


The resulting scatterplot has two defining features:

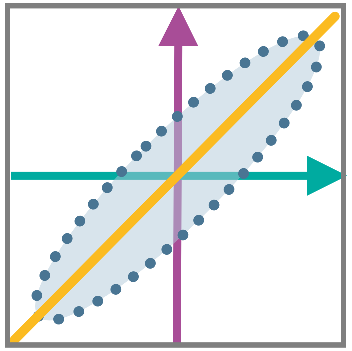
- The **spread** around the trendline (quantified by r-Alerting)
- The **slope** of the trendline

Comparative Attenuation

Correlation of correlations



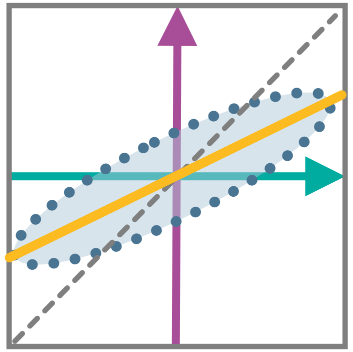
Low r-Alerting
Wide spread
Slope ≈ 1



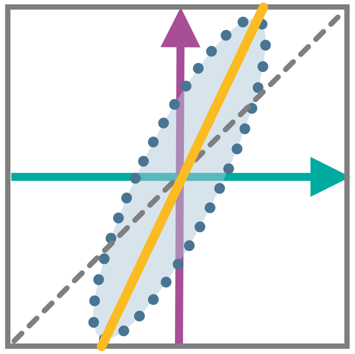
High r-Alerting
Good linear fit
Slope ≈ 1

High r-Alerting (~good linear fit)
implies **good conceptual comparability**
However, the **slope** should also be **close to 1!**

Slopes

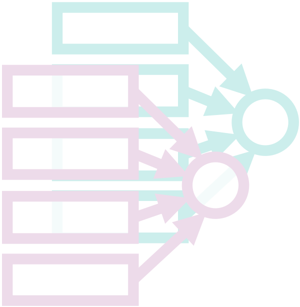


High r-Alerting
Good linear fit
Slope < 1



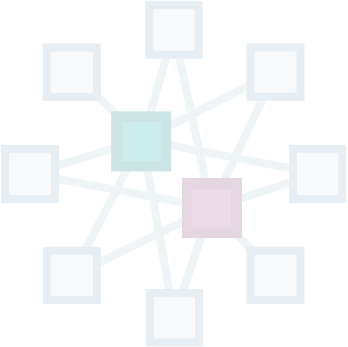
High r-Alerting
Good linear fit
Slope > 1

However, **good linear fit but a slope ≠ 1**
might imply a **global difference in random errors**
between the modes!



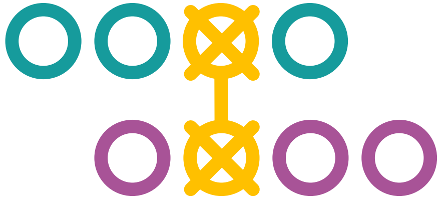
Formal Measurement Invariance

MGCFA



Concepts and Reliability

R-Alerting and comparative attenuation



Aligning measurement units

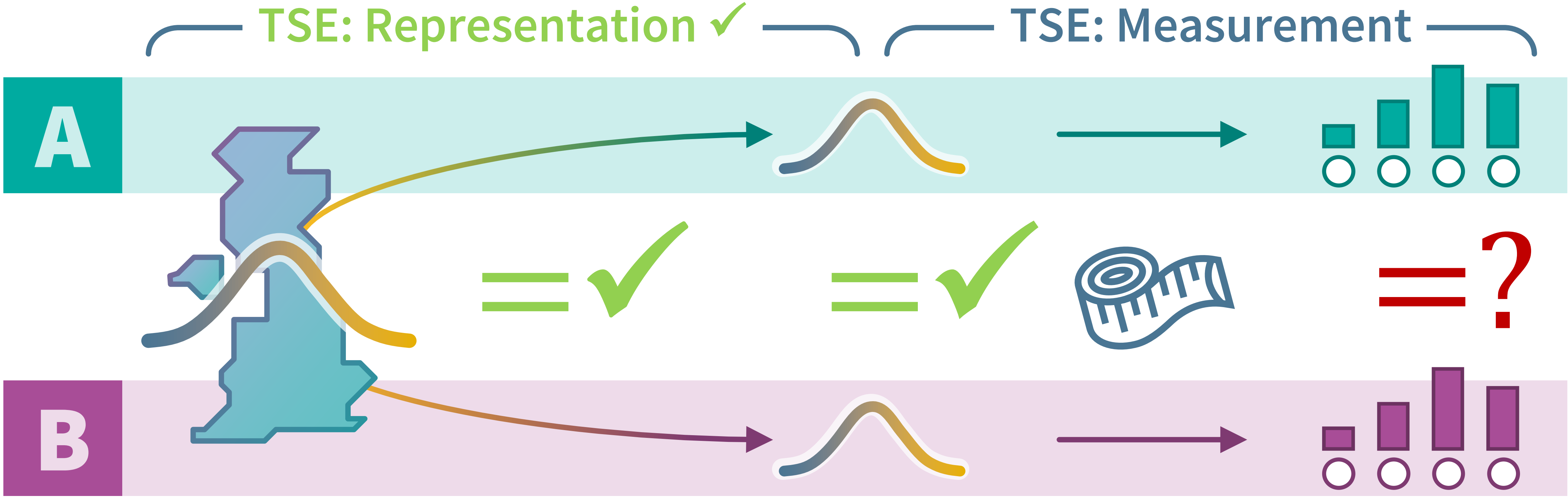
OSE-RG

Generalizable Mode Effects

MTMM Meta-Analysis with SQP

gesis Leibniz Institute for the Social Sciences

Random Groups Design
(= random experiment for Equating)



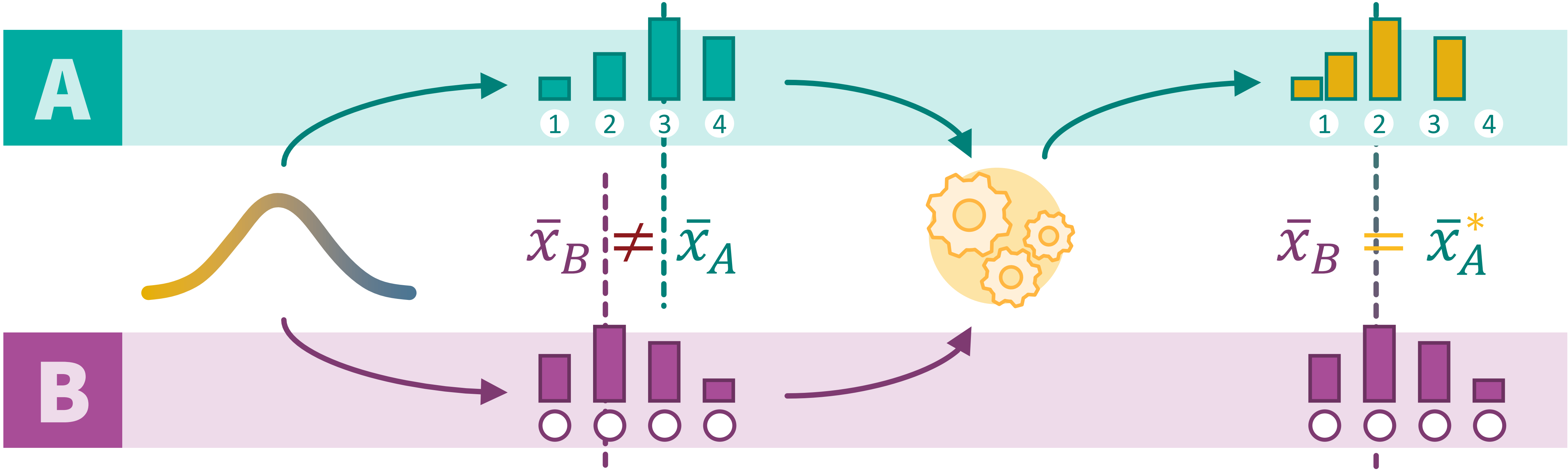
Equally good
Random samples
in modes **A** and **B**

Identical
Latent
distributions
in both samples

Distribution Differences
=
Measurement Unit Differences

OSE-RG: Observed-Score Equating

in a Random Groups Design



Identical Latent distributions in both samples

Different Response Distributions

Align Response Distributions with **OSE-RG**

Recoded mode **A** now **fits** mode **B**

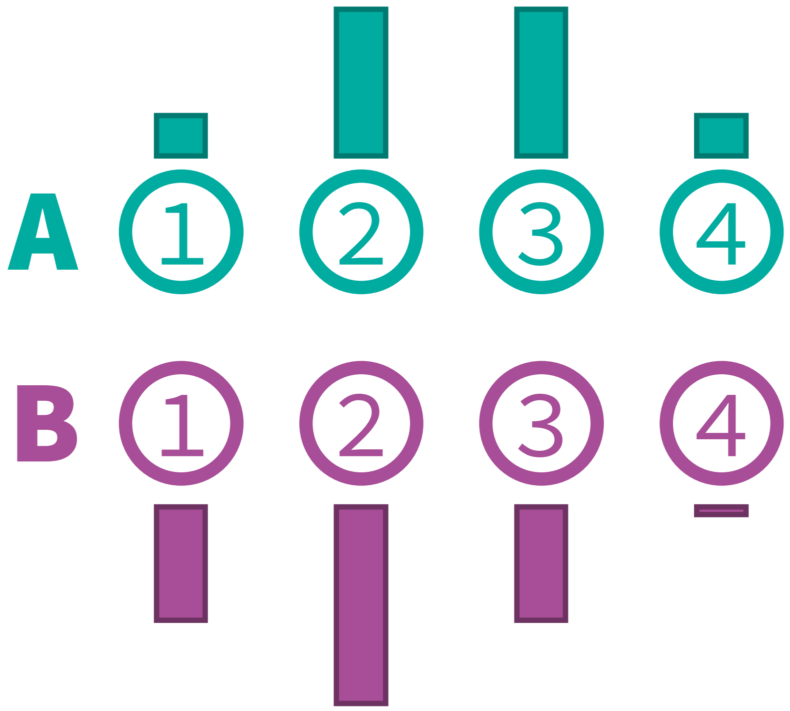
Linear Equating Algorithm: Recoding A to B

Response distributions

for A and B

in a **random groups design**

Differences in distribution shape are measurement differences, not true differences



Linear Equating Algorithm: Recoding **A** to **B**

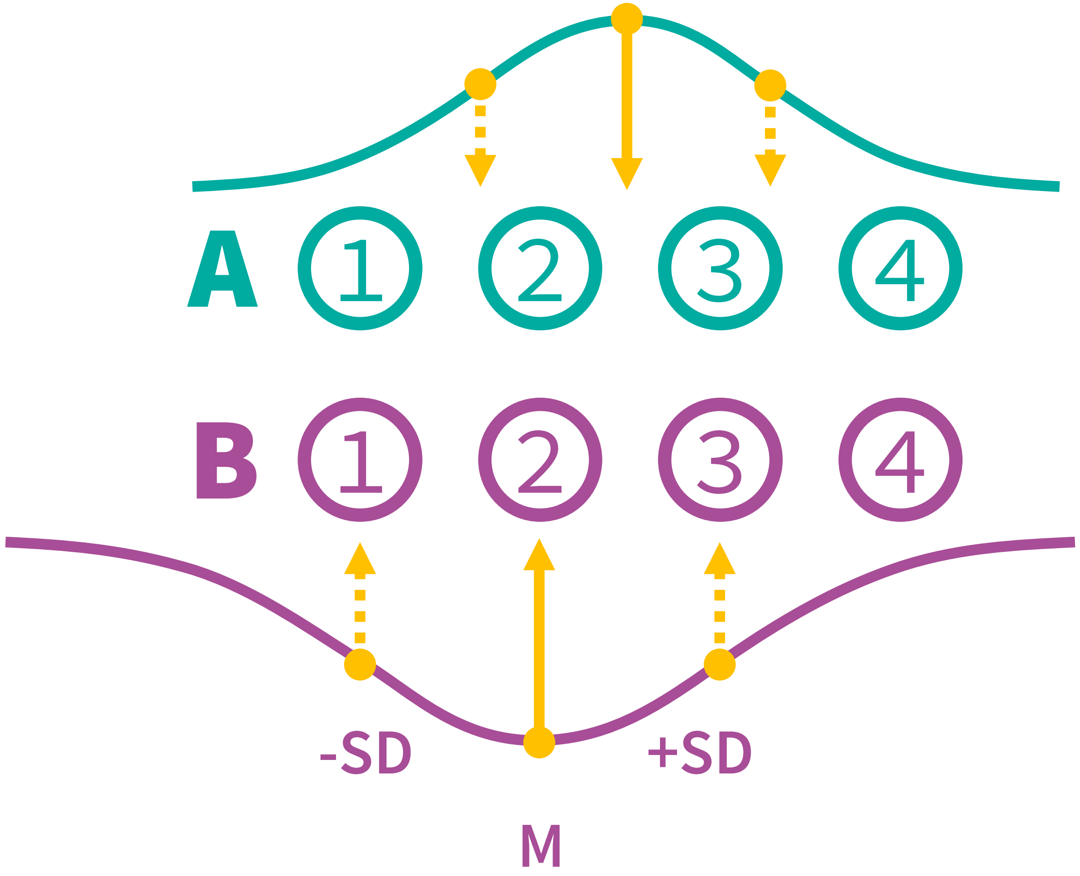
Response distributions

for A and B

in a random groups design

…

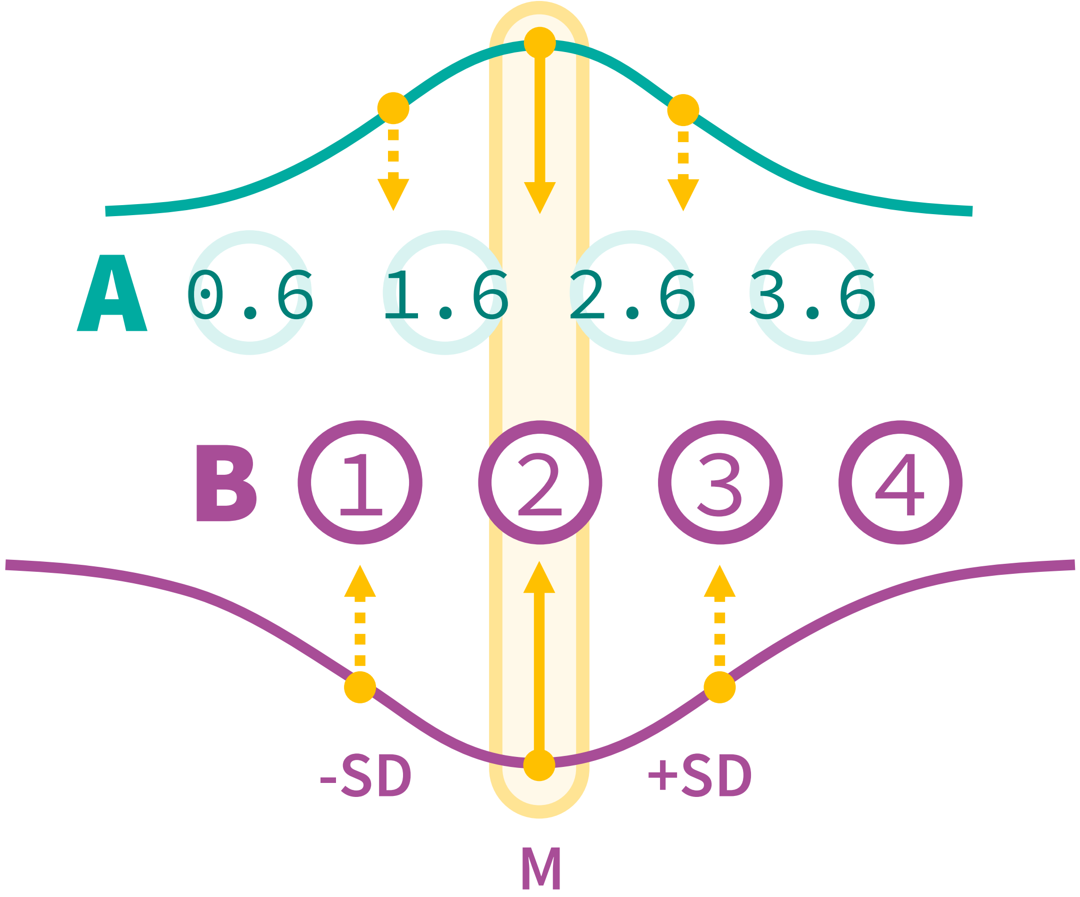
simplified to two parameters
Mean and **SD**



Linear Equating Algorithm: Recoding A to B

Linear transformation to **recode scores** of A towards the **measurement scale** of B…

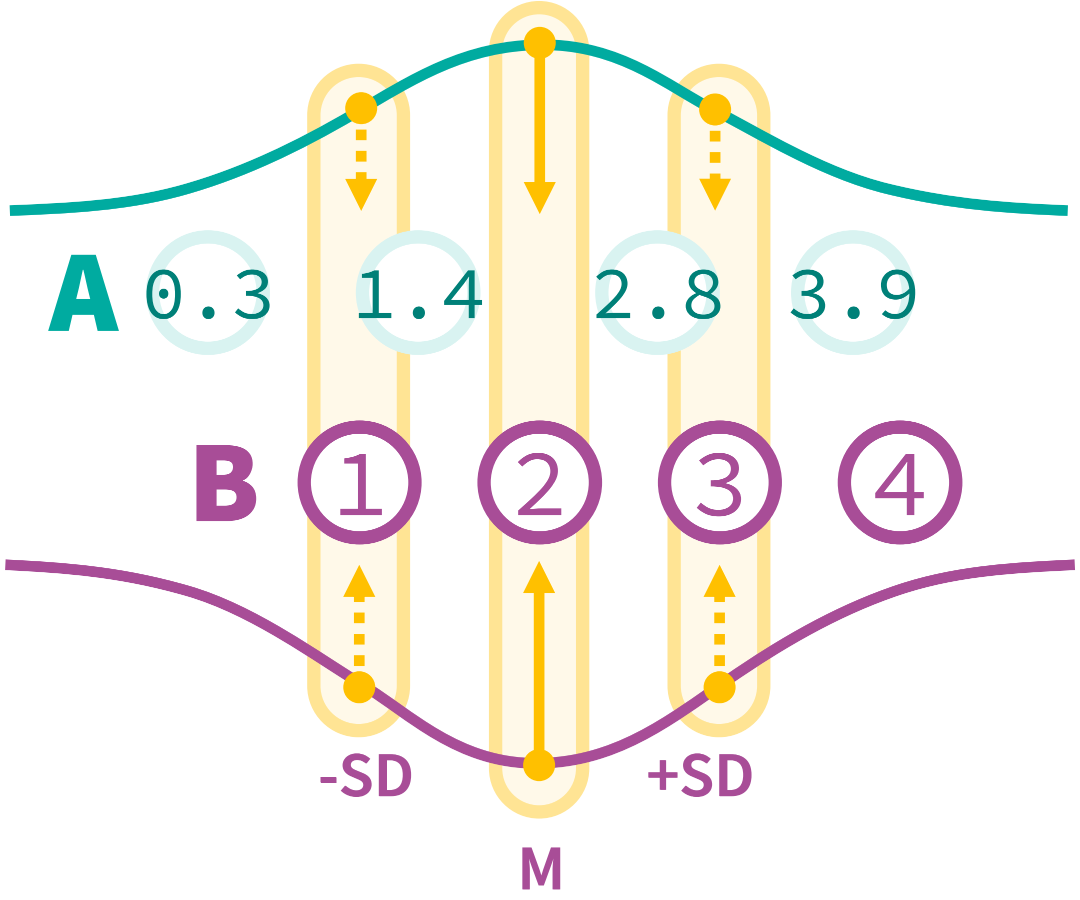
1. Aligning the **means**



Linear Equating Algorithm: Recoding A to B

Linear transformation to **recode scores** of A towards the **measurement scale** of B…

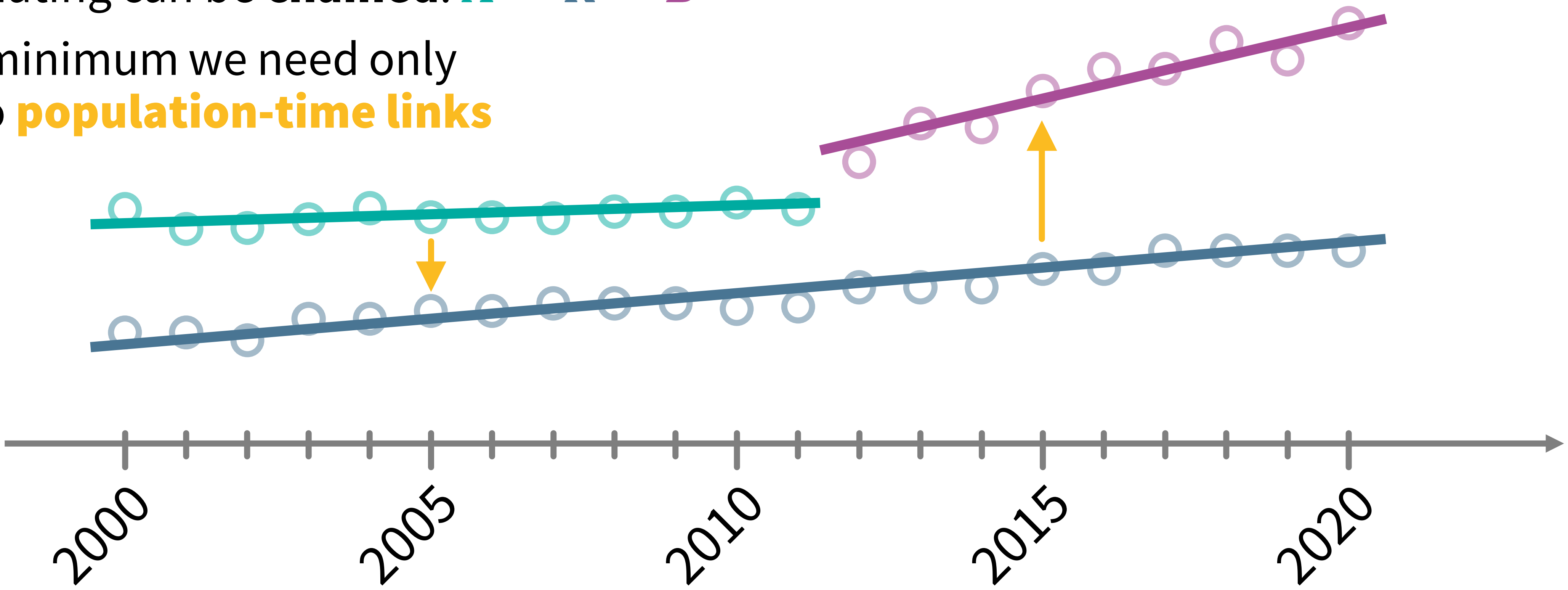
1. Aligning the **means**
2. and the **standard deviations**



OSE-RG with a reference survey program

(with probabilistic samples of the same population)

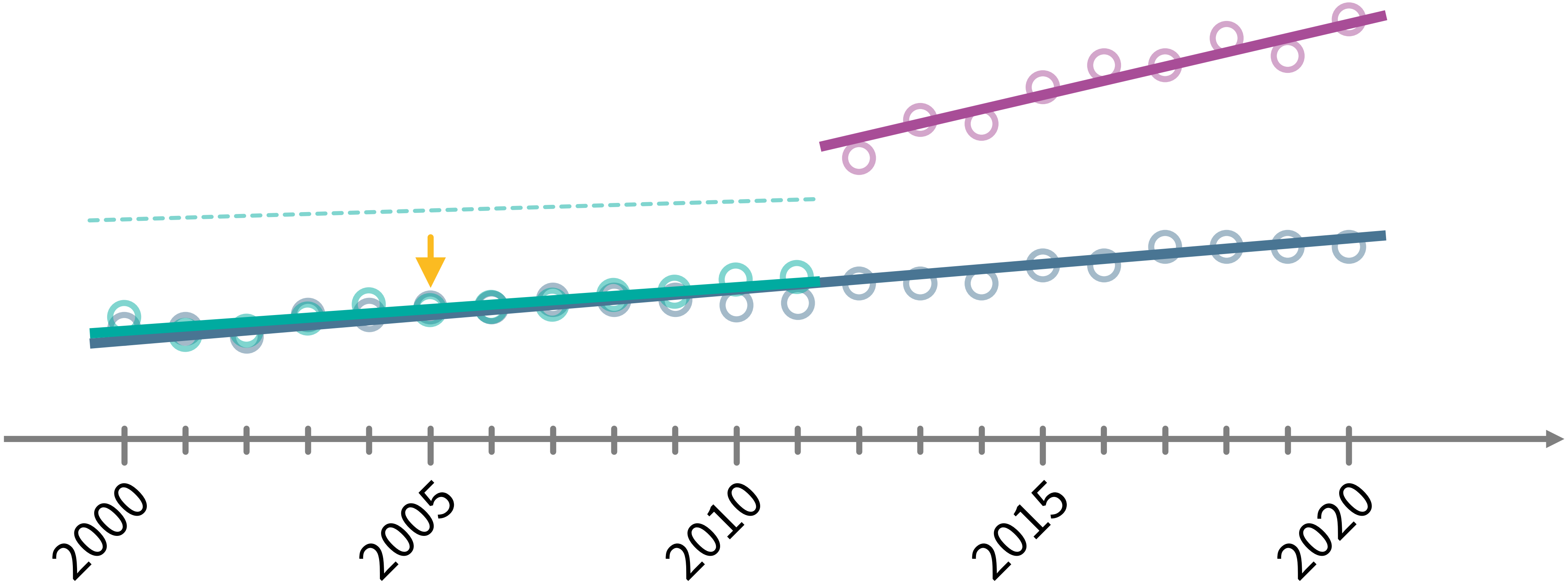
- **Two surveys randomly sampling the same country in the same year** are also a **random groups design**!
- Equating can be **chained**: **A** → **R** → **B**
- At minimum we need only two **population-time links**



OSE-RG with a reference survey program

(with probabilistic samples of the same population)

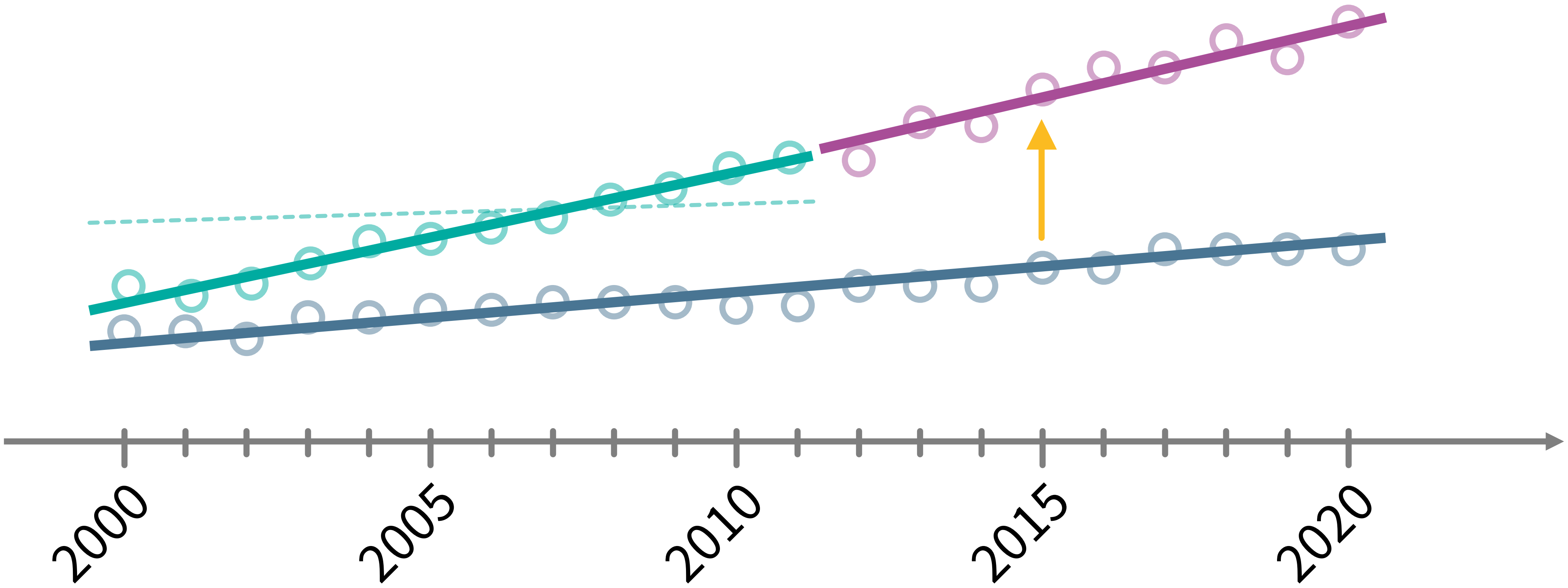
A → R → B



OSE-RG with a reference survey program

(with probabilistic samples of the same population)

A → R → B



Observed-Score Equating
in a Random Groups Design

Points to consider:

- OSE-RG only **aligns Measurement Units**
- **Systematic** and **random measurement errors** are preserved
- **Mode dependent errors of representation**
  can bias the Equating Result!

Mitigating differences in representation:

- **Adjustment weights**
- **NEC Equating** (Non-equivalent groups with covariates design)

Formal Measurement Invariance

MGCFA

Concepts and Reliability

R-Alerting and comparative attenuation

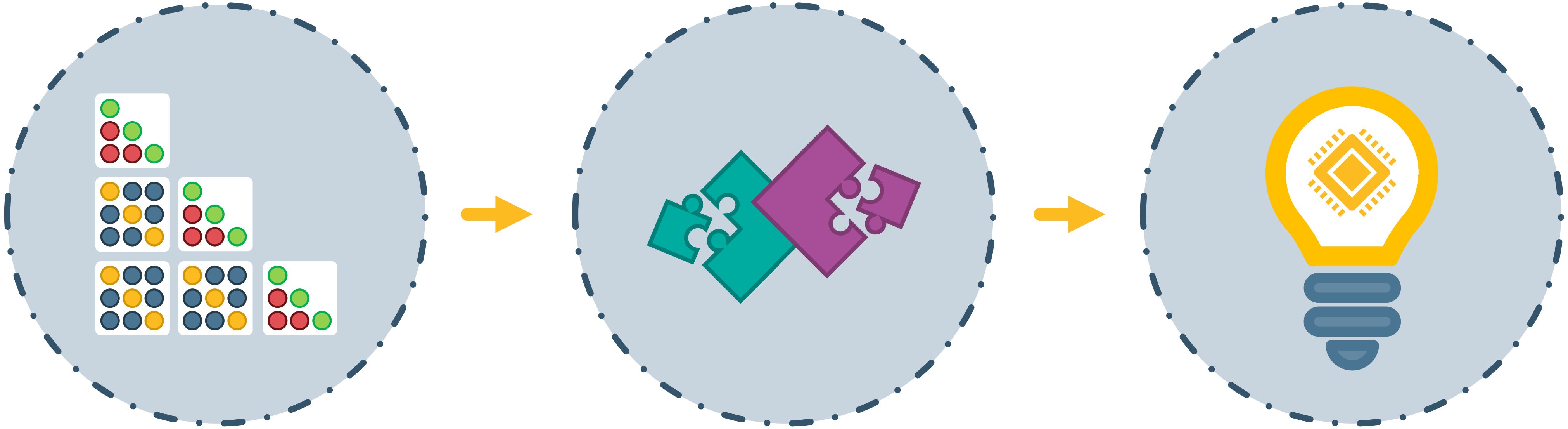
Aligning measurement units

OSE-RG

Generalizable Mode Effects

MTMM Meta-Analysis with SQP

Primer on the SQP 3.0 Survey Quality Predictor



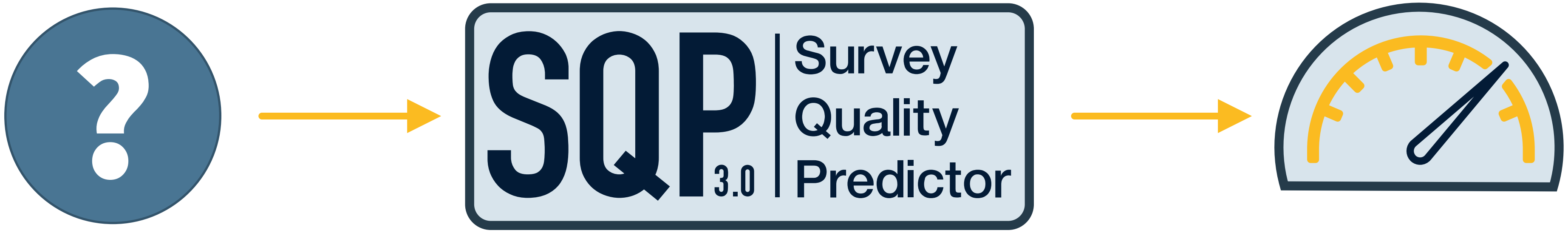
ESS

MTMM Experiments
evaluating the measurement quality of >6000 instruments in 33 countries

Coding
a set of formal design characteristics

Meta-Analysis
predicting measurement quality via these characteristics

SQP for users



Coding
the formal characteristics of a question to be evaluated

SQP
determines the likely quality based on the meta-analysis

Quality estimates
are given as point estimates with ranges

SQP in survey mode harmonization

SQP has several **characteristics** of interest for **survey mode** harmonization!

- Showcards or other **visual aid** used?
  - **Horizontal** or **vertical scale**?
  - …
- **Computer assisted** answer registration?
- **Interviewer** or **self-completion**?
- **Visual** or **oral** presentation?

SQP in survey mode harmonization

Quality Estimates

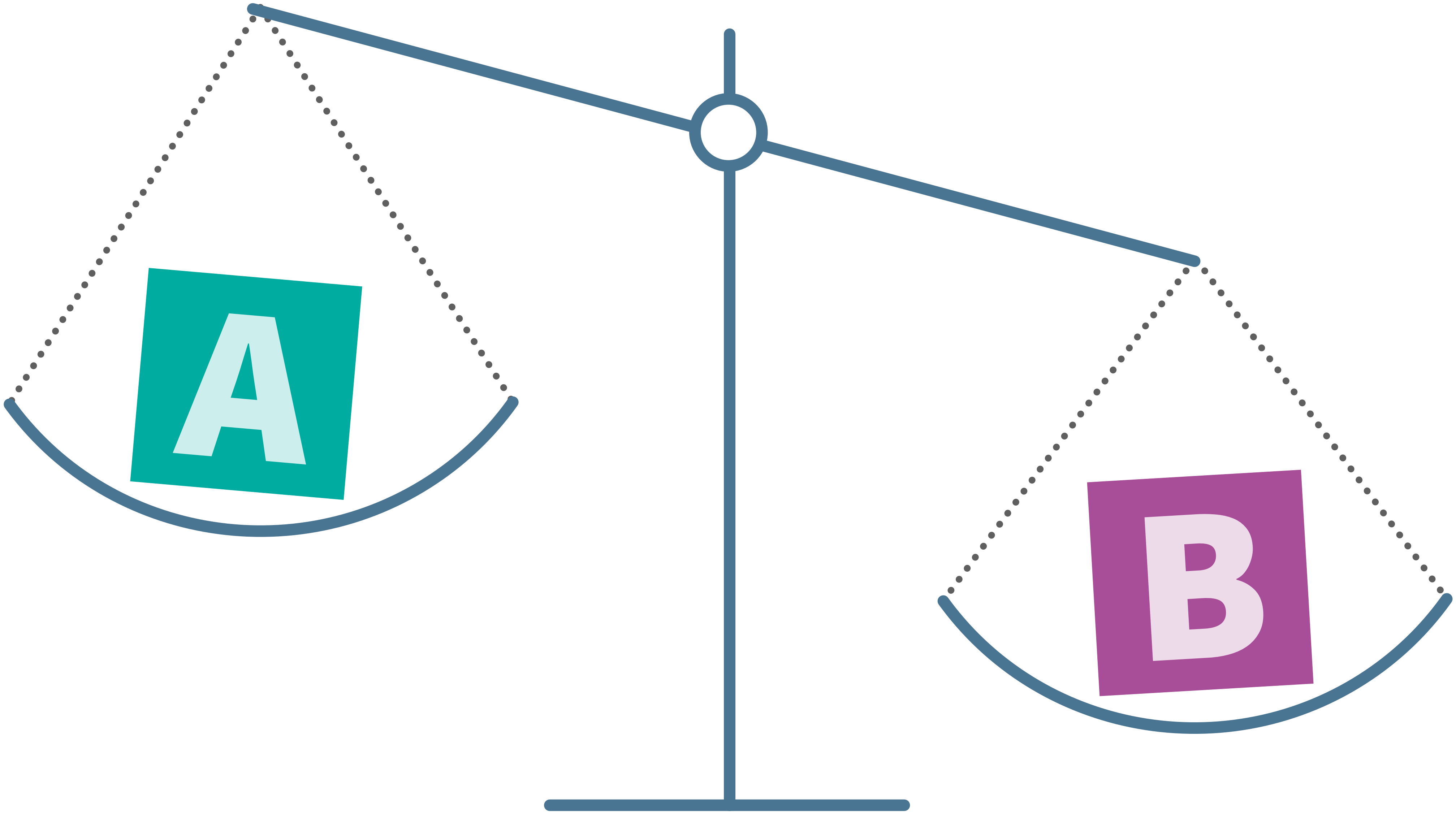
Predicting the quality of indivisual questions in both modes

Generalizable Effects?

Querying the meta-analysis for general effects of mode relevant charateristics

Meta-Analytical Framework

Adding new MTMM-Mode Experiments to the SQP Pool


A
B

Generalizability across…?

Specific Questions

Modes may have very different effects on different instruments

Countries / Cultures

Modes may have different effects in different countries / cultures / languages

Respondents

Specific respondents or specific suppopulations may react differently to different modes

However, searching for generalizable methodological differences between modes is still important!

Healthy Pragmatism

- Modes **can** matter, but they **do not have to** matter

- **Comparability** brings methodological issues into **sharp contrast**. However, we should not be stricter in comparability than we are in single-mode data

- **Quantifying issues** is often all it takes to **mitigate issues**

Ressources

GESIS Blog Series on (Instrument) Harmonization
https://blog.gesis.org/adventures-in-ex-post-harmonization-frankensteins-creature/

SQP 3.0
https://sqp.gesis.org/

GESIS consultation on harmonization
https://www.gesis.org/en/services/crm/request-form-for-consultations-and-scientific-services

Singh, R. K. (in print). Harmonizing single-question instruments for latent constructs with equating using political interest as an example. *Survey Research Methods*

Thank you for your attention!

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz Leibniz-Gemeinschaft