www.surfingthegulf.com

Questionnaires and survey data collection

Jakub Procházka

DXH_MET1: Methodology 1 (2023)

Let's try a questionnaire
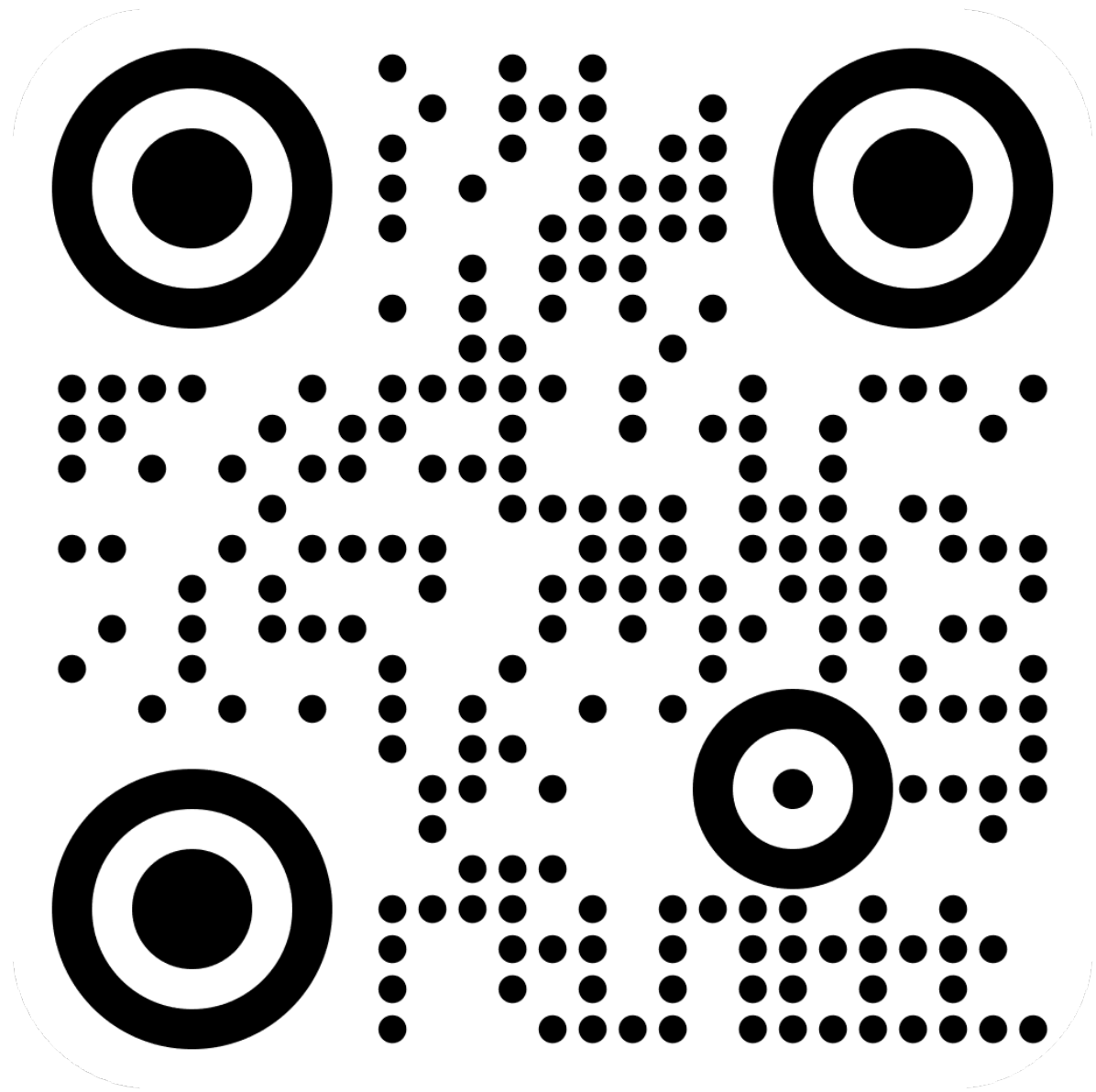
H. Cígler, K. Rečka a M. Tancoš (FSS MU): Brno height questionnaire

http://fssvm6.fss.muni.cz/vyska/

MEASUREMENT MODEL

Error

$$X = T + E(r) + B + E(w)$$

Observed Score

True Score

Random Error

Bias

Error of Wrong Construct

Unsystematic Chance Factor

Non-Random Error

Systematic

Reliability

Validity

Picture: Chitvan Trivedi, https://conceptshacked.com/measurement-model/
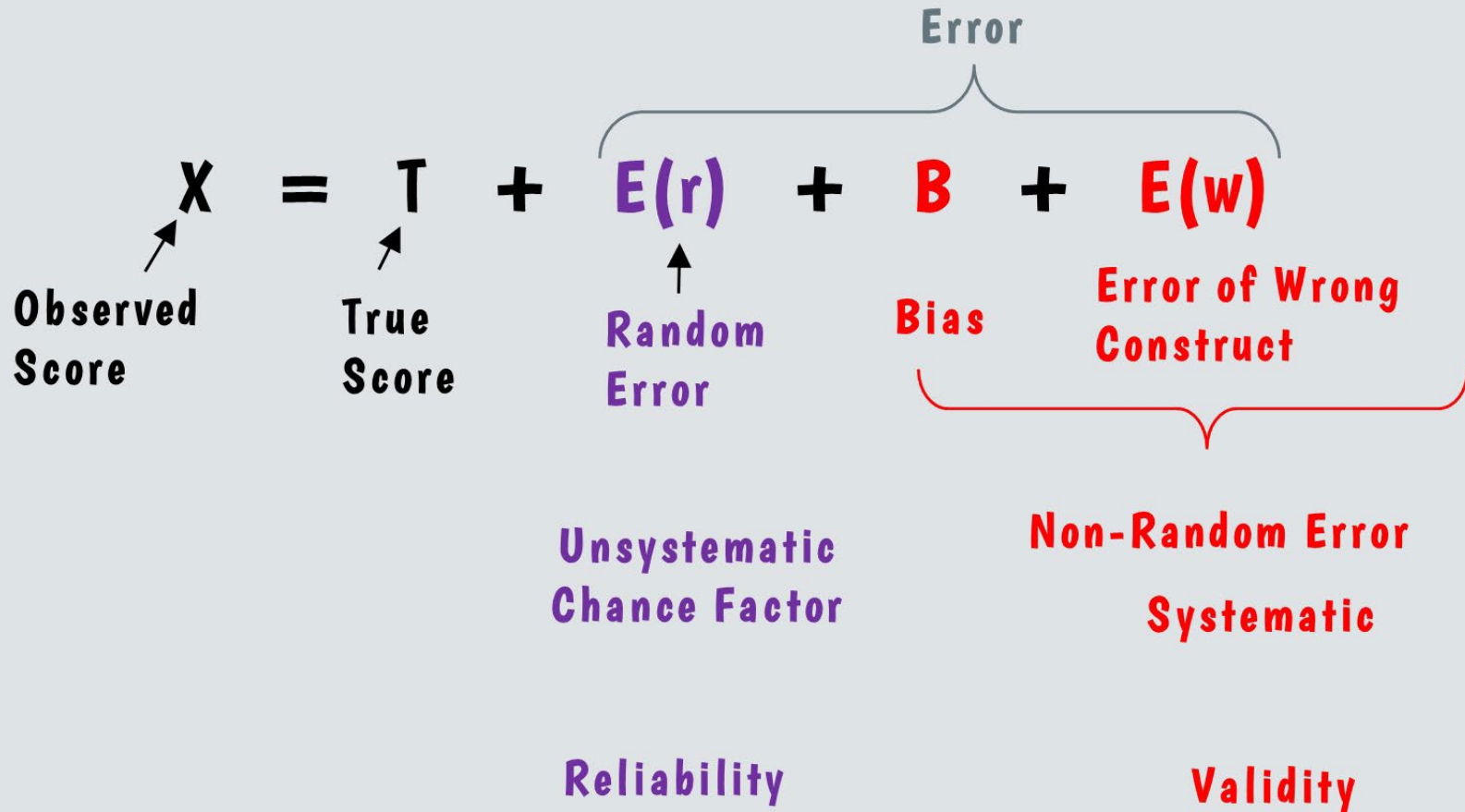
Reliability

- "Spolehlivost" in Czech
- How consistent the results provided by the instrument are in the conditions where they should be consistent.

If I measure Peter's height by repeatedly attaching this ruler:
Test-retest reliability:
- Will I get the same result every time with 10 measurements?

Inter-rater reliability:
-Will I get the same result as Kate and John if they measure Peter's height with the same ruler?

Split-half reliability:
-Will I get the same result if I measure with the first half of the ruler and the second half of the ruler?

Picture: Michal Kalaš
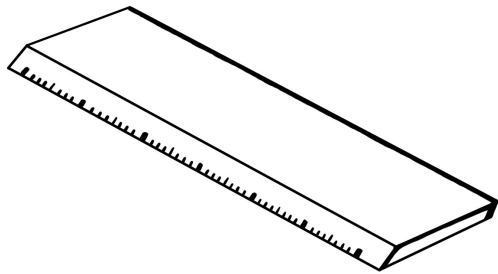
Reliability estimations

Test-retest reliability:
- When I measure the same thing with the same measurement tool over time, the results correlate strongly with each other.

Inter-rater reliability:
- When multiple people measure the same thing with the same measuring instrument, the results are strongly correlated with each other.

Split-half reliability:
- Results measured by two parts of the same method are strongly correlated with each other.

Internal consistency
- The total variance of a measurement instrument is largely explained by the shared variance of its subparts (simply e.g. for a questionnaire: items are strongly correlated with each other)

Measured using Cronbach's alpha, McDonald's omega etc.

Example: http://fssvm6.fss.muni.cz/height/

Validity

- "Platnost" in Czech
- To what extent does the instrument measure what it is supposed to measure

If I measure the height of 1,000 people by repeatedly attaching this ruler:

Content validity:
- Would such a measurement be consistent with the theory of how height should be measured?

Construct (convergent) validity:
- Will the height measured by the ruler correlate moderately with participants' weight?

Criterion-related (concurrent) validity:
- Will the outcome correlate strongly with an outcome of a certified platinum-iridium ruler?

Criterion-related (predictive) validity:
- Will the result allow me to predict who will bang their head on the door frame?

Picture: Michal Kalaš
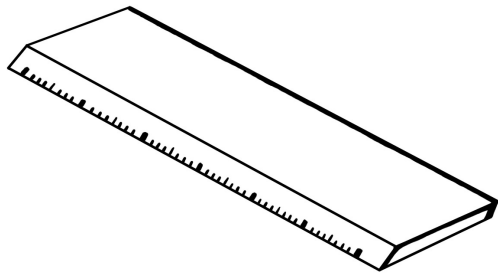
Validity estimation

Content validity:

- The degree to which the content of the test and the way how the construct is measured correspond to how the construct is defined according to the theory.

- Experts agree that a method measures what it is intended to measure.

Example: a questionnaire used to assess an employee's performance lists only performance-related items and does not miss any essential component of performance.

Validity estimation

Construct validity:

- Convergent validity: The degree to which measures of two constructs that should be related to each other according to theory are related.

- Divergent validity: The extent to which measures of two constructs that should not be related according to theory are unrelated.

Example: the results of a questionnaire measuring task job performance are related to supervisor's satisfaction with the employee but are unrelated to the results of a questionnaire measuring an employee's extraversion.

- Factor(ial) validity: the degree to which the covariance of measured items matches the real covariance or behaviors in real life.

Example: The confirmatory analysis shows, that the data gathered by the job performance questionnaire with 3 subscales correspond to the theoretical 3-factor model of job performance.

Validity estimation

Criterion-related validity:

- The degree to which a measurement result is related to a criterion that well represents the construct being measured.

- Concurrent validity: The degree to which a measurement result is related to another measurement result (some standardised indicator) applied at the same time.

Example: the results of a job performance questionnaire completed by a supervisor are strongly correlated with KPIs evaluation.

- Predictive validity: The degree to which a measurement result is related to a criterion observed in the future.

Example: sales skills test scores are strongly related to the number of new orders won by sales reps in the following year.

Reliability and validity

- The method must have sufficient reliability and validity to be trusted.

- A method with low reliability cannot be valid.

Example: I want to measure job performance using a crystal ball. Different fortune tellers using the same ball will arrive at different results (low reliability). Such a measurement will probably not be valid (low validity).

- A method with high reliability may not be valid.

Example: I measure job performance of sales representatives by measuring their height by a certified platinum-iridium ruler. I measure height very reliably (high reliability), but the performance measurement is probably not valid (low validity) because physical height is not very relevant for sales.

→ **I need to consider the validity and reliability of all questionnaires which I want to use.**

→ **I need to be able to provide evidence about the reliability and validity of questionnaires during the review process.**

Reliability and validity
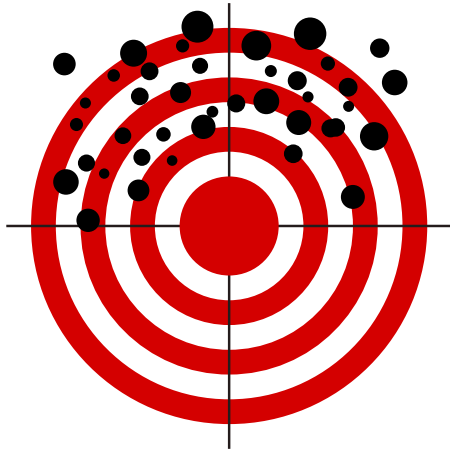
Not reliable and
not valid

Not reliable and
therefore not valid
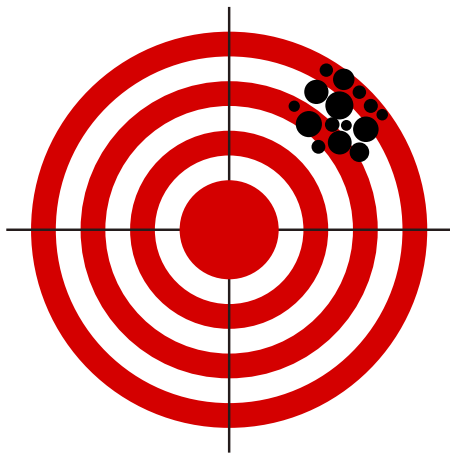
Reliable but not
valid

Reliable and valid

Picture: Nevit Dilmen
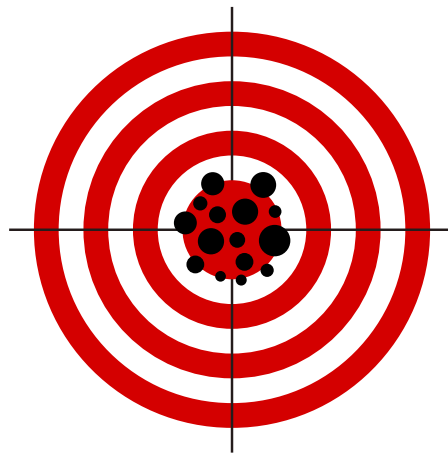
How to get a reliable and valid questionnaire

- Use an existing questionnaire
  - Easiest option
  - Availability of evidence about validity and reliability from past research
  - Need to provide evidence about reliability and validity for a specific population
  - Possibility to compare results with prior research
  - May not meet the needs of research

- Adapt an existing questionnaire
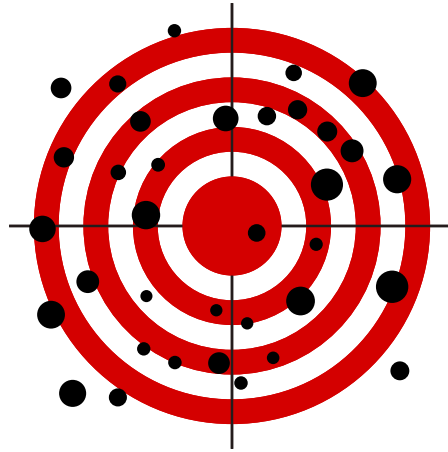 - Adaptation to a new language and/or context (type of organization, time frame, culture…)
  - Need to demonstrate equivalence (for cross-cultural comparison, for using existing evidence about validity)
  - Responsibility to provide evidence about the reliability and validity of the adapted version
  - More than 30 guidelines on how to create a new language version of a questionnaire

- Create your own questionnaire
  - Hardest option
  - Potential problems with content validity
  - Great reviewer attention is paid to the new questionnaires
  - Many guidelines on how to create a questionnaire

Questionnaire development
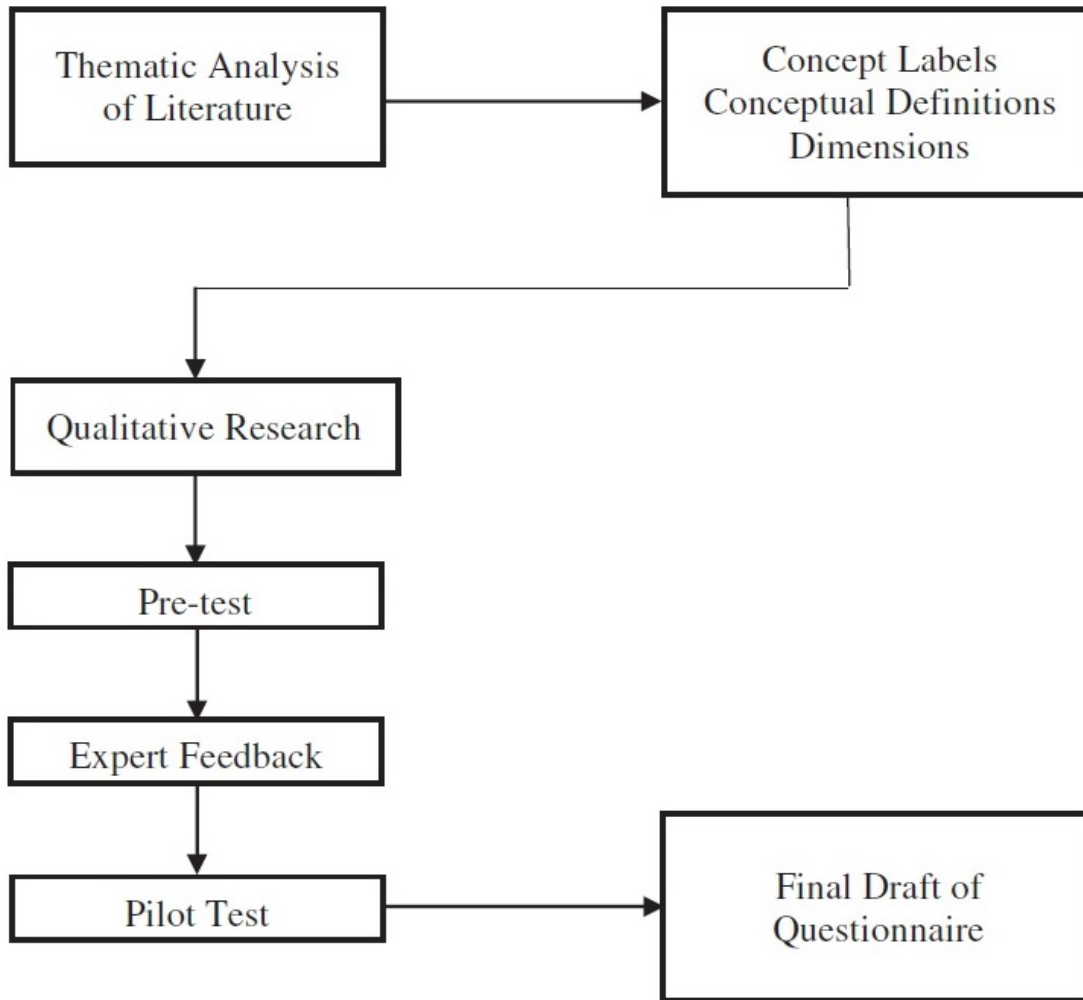
Figures sources: Carpenter (2018) and Boateng et al. (2018)
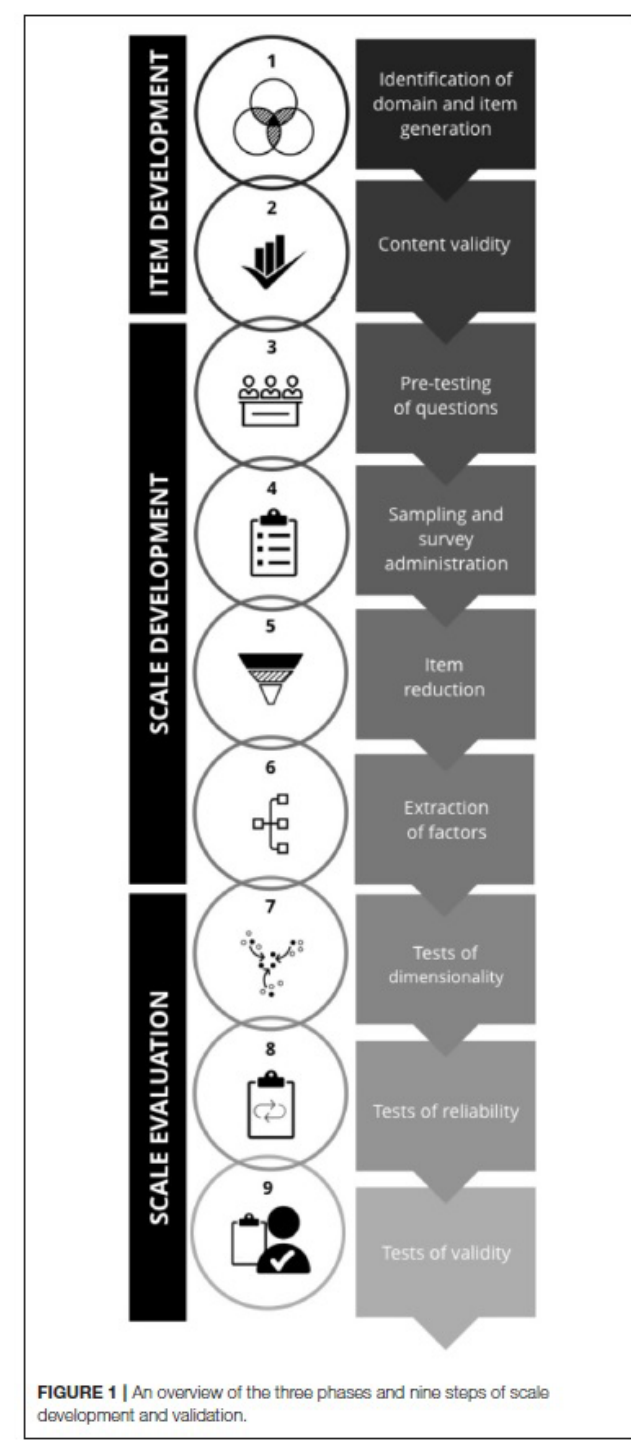
FIGURE 1 | An overview of the three phases and nine steps of scale development and validation.

Questionnaire development

1. Literature review
Definition of construct/s
or description of domain

2. Item development
Deductive (from definition to item)
or inductive (to describe complete domain)

3. Qualitative item reduction + rephrasing
Unclear, irrelevant, recurring items etc.
Research team + cognitive interviews

4. Quantitative item reduction (pre-test)
Pilot study with dozens of respondents
Internal consistence, variability, feedback…

5. Establishing content validity
Expert feedback
Content validity ratio, Q-sorting etc.

6. Quantitative pilot study
Pilot study with hundreds of respondents
Construct & criterion validity, reliability

7. Final item reduction (if needed)
Unclear, irrelevant, recurring items etc.
Research team + individual respondents

8. Validation study
Evidence on validity and reliability of final
questionnaire

See Boateng et al. (2018), Hinkin (1998) and Crawford & Kelder (2019).

Qualitative item reduction: Focus on distorted questions

1. Problematic wording
Ambiguous, double-barreled
Hard-to-understand, too complex…

2. Response scale problems
Too short or too long, forced choice, vague
Missing or overlapping intervals…

3. Captures inadequate data
Categories instead of open answer
Hypothetical question…

4. High risk of biased answer
Leading question, social desirable answer
Framing, demanding on memory…

See Tourangeau et al. (2000)
and Choi & Park (2005)

Questionnaire adaptation

Table 1. Possible Scenarios Where Some Form of Cross-Cultural Adaptation is Required

Wanting to use a questionnaire in a new population described as follows:	Results in a Change in...			Adaptation Required	
	Culture	Language	Country of Use	Translation	Cultural Adaptation
A Use in same population. No change in culture, language, or country from source	—	—	—	—	—
B Use in established immigrants in source country	✓	—	—	—	✓
C Use in other country, same language	✓	—	✓	—	✓
D Use in new immigrants, not English-speaking, but in same source country	✓	✓	—	✓	✓
E Use in another country and another language	✓	✓	✓	✓	✓

Adapted from Guillemin et al.[4]

Beaton et al. (2000)

See Beaton et al. (2000) and Epstein et al. (2015)

Questionnaire adaptation

Stage I: Translation
- Two translations (T1 & T2)
- into target language
- informed + uninformed translator
→ *written report for each version (T1 & T2)*

Stage II: Synthesis
- synthesize T1 & T2 into T-12
- resolve any discrepencies with translators' reports
→ *written report*

Not nessesary
(Epstein et al. (2015)

Stage III: Back translation
- two english first-language
- naive to outcome measurement
- work from T-12 version
- create 2 back translations BT1 & BT2
→ *written report for each version (BT1 & BT2)*

Stage IV: Expert committee review
- Review all reports
- methodologist, developer, language professional, translators
- reach consensus on discrepencies
- produce Pre-final version
→ *written report*

Stage V: Pretesting
- n=30-40
- complete questionnaire
- probe to get at understanding of item
→ *written report*

Stage VI: Submission and Appraisal of all written reports by developers/committee

Beaton et al. (2000)

Main issues with survey data collection

1. Low quality instruments
Low reliability or validity
Inequivalent adaptations...

2. Sampling problems
Non-representative sample
Low response rate (non-response bias)...

3. Inattentive respondents
Low motivation of respondents (incentives?)
Long questionnaire...

4. Biased answers
Low anonymity, context of data collection
Order of questionnaires, social desirability...

5. Common-method bias
Systematic error variance shared among
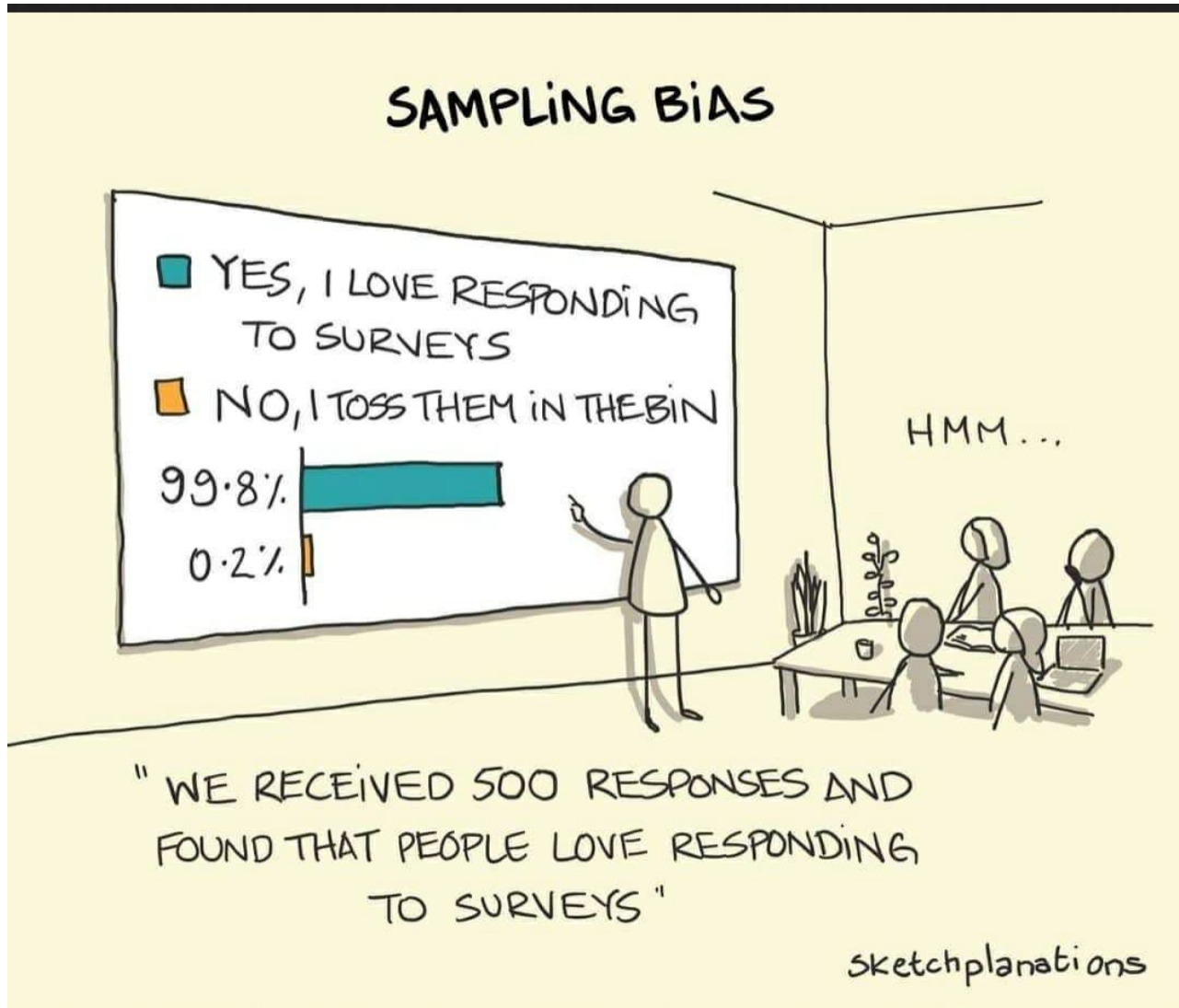variables measured by the same way

6. Data fishing (dredging) in large surveys
Multiple predictors, DV, analyses, p-hacking...
Solution: pre-registration

See Podsakoff (2003) for more details about common-method bias
See Erasmus et al. (2022) for more details about data fishing

Issue: Sampling problems

Issue: Dealing with inattentive respondents

Measuring (page / questionnaire) response time

- Comparing response time to the rest of the sample or some standard

Attention checks

- Specific questions within the survey: "To monitor quality, please respond with a two for this item."

Response consistency analysis

- Post-hoc analysis: Consistent responses to similar questions

Multivariate outlier analysis

- Post-hoc statistical analyses of outliers

Self-report diligence

- Special question at the end of the questionnaire
- "I carefully read every survey item."

Identified (not anonymous) answers

- May cause ethical problems and bias connected to social desiability

See Buchanan & Scofield (2018) and Meade & Criag (2012)

Thank you for your attention…

References

- Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, *25*(24), 3186-3191.

- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in Public Health*, *6*, 149.

- Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*, *50*(6), 2586-2596.

- Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures*, *12*(1), 25-44.

- Choi, B. C., & Pak, A. W. (2005). A catalog of biases in questionnaires. *Preventing Chronic Disease*, *2*(1), PMC1323316.

- Crawford, J. A., & Kelder, J. A. (2019). Do we measure leadership effectively? Articulating and evaluating scale development psychometrics for best practice. *The Leadership Quarterly*, *30*(1), 133-144.

- Epstein, J., Santo, R. M., & Guillemin, F. (2015). A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. *Journal of Clinical Epidemiology, 68*(4), 435-441.

- Erasmus, A., Holman, B., & Ioannidis, J. P. (2022). Data-dredging bias. *BMJ Evidence-Based Medicine*, *27*(4), 209-211.

- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437-456.

- Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *885*(879), 1010-1037.

- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.