Shaken or stirred?

Mixing methods in evaluation

Prof. Dr. Jan Hense

INTOSAI Working Group on Evaluation of Public Policies and Programs Forum 2022

Bern, Switzerland

28 September 2022

My name is Hense – Jan Hense

- Training in psychology and education
- 22 years in evaluation as evaluator, teacher, researcher, and consultant
- Evaluation at local, regional, national, and EU levels
- 20 years in the academia, 5 as full professor
- Independant evaluation consultant since 2020
- Board member/president of the Gesellschaft für Evaluation – DeGEval from 2015 to 2021

Before we start: What is your background re evaluation?

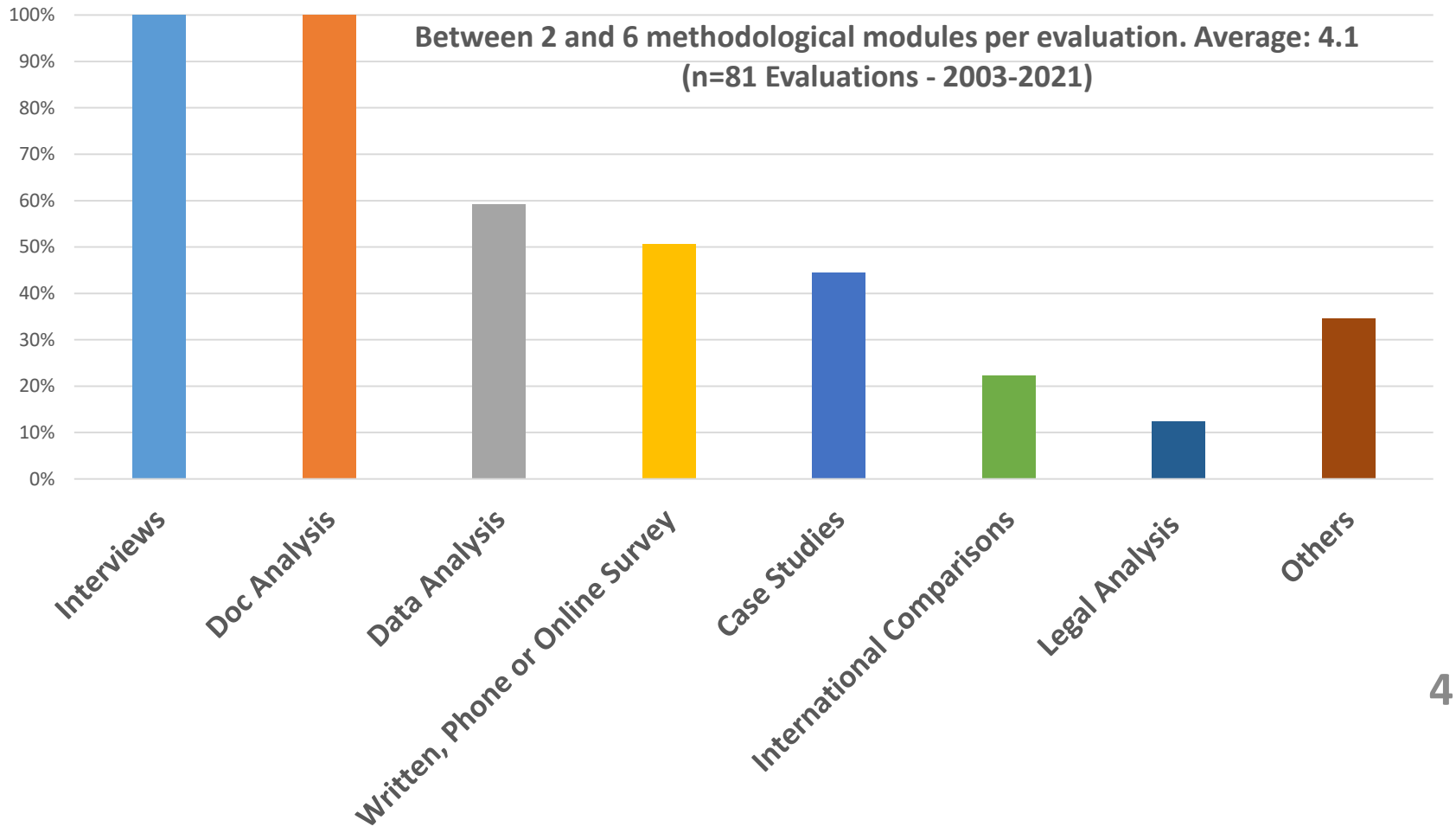
- Role:
  - Commissioner of evaluation studies
  - Consumer of evaluation studies
  - Evaluator
  - Other

- Role of formative / summative evaluation?
  - Summative: evaluation for decision and accountability
  - Formative: evaluation for improvement and learning

- Key terms:
  - Quantitative/qualitative methods
  - Program theory (theory of change, logic models etc.)
  - Causation

Mixing methods at the SFAO

Used Methods per Evaluation at the SFAO



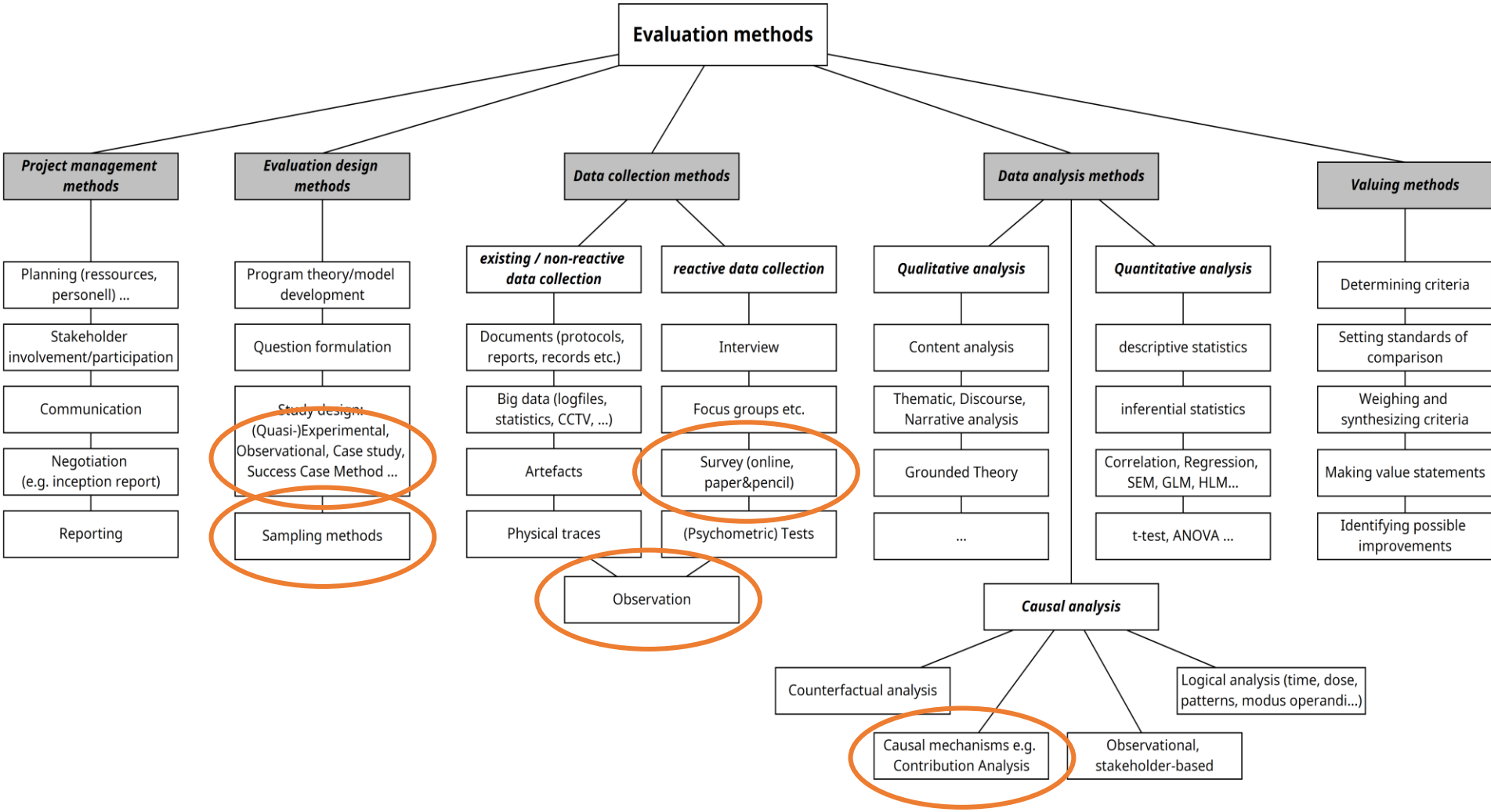
Topics

1. List of ingredients: Evaluation methods
2. The importance of sampling
3. Surveys: the good, the bad, and the ugly
4. Less common methods:
    1. Observational methods
    2. Case studies
    3. Success case method
    4. Contribution analysis
5. How to mix the perfect Martini

The full list of ingredients: Evaluation methods

Evaluation „methods“

- Data collection methods
  - Gathering data
- Data analysis methods
  - Making sense of data, answering questions
- Project management methods
  - Planning and implementing an evaluation
- Evaluation design methods
  - Setting the frame for reaching evaluation goals
- Valuing methods
  - Deriving value judgments


Evaluation methods
Project management methods
Planning (ressources, personell) ...
Stakeholder involvement/participation
Communication
Negotiation (e.g. inception report)
Reporting
Evaluation design methods
Program theory/model development
Question formulation
Study design: (Quasi-)Experimental, Observational, Case study, Success Case Method ...
Sampling methods
Data collection methods
existing / non-reactive data collection
Documents (protocols, reports, records etc.)
Big data (logfiles, statistics, CCTV, ...)
Artefacts
Physical traces
reactive data collection
Interview
Focus groups etc.
Survey (online, paper&pencil)
(Psychometric) Tests
Observation
Data analysis methods
Qualitative analysis
Content analysis
Thematic, Discourse, Narrative analysis
Grounded Theory
...
Quantitative analysis
descriptive statistics
inferential statistics
Correlation, Regression, SEM, GLM, HLM...
t-test, ANOVA ...
Causal analysis
Counterfactual analysis
Causal mechanisms e.g. Contribution Analysis
Observational, stakeholder-based
Logical analysis (time, dose, patterns, modus operandi...)
Valuing methods
Determining criteria
Setting standards of comparison
Weighing and synthesizing criteria
Making value statements
Identifying possible improvements

Sampling

Population and sample

Population = all subjects of interest

Sample = selection of subjects from the population

- Sample is always smaller
- Sample is used to represent the population *for practical reasons*

Examples

Population	*Sample*
• Wine vintage	• Glass of wine
• Water in a lake	• Bottle of water collected from lake
• Electorate in a country	• Election poll
• Readers of a newspaper	• Contributors to the „letters to the editor“
• Participants of a training measure	• Participants present at the last meeting

What makes a good sample?

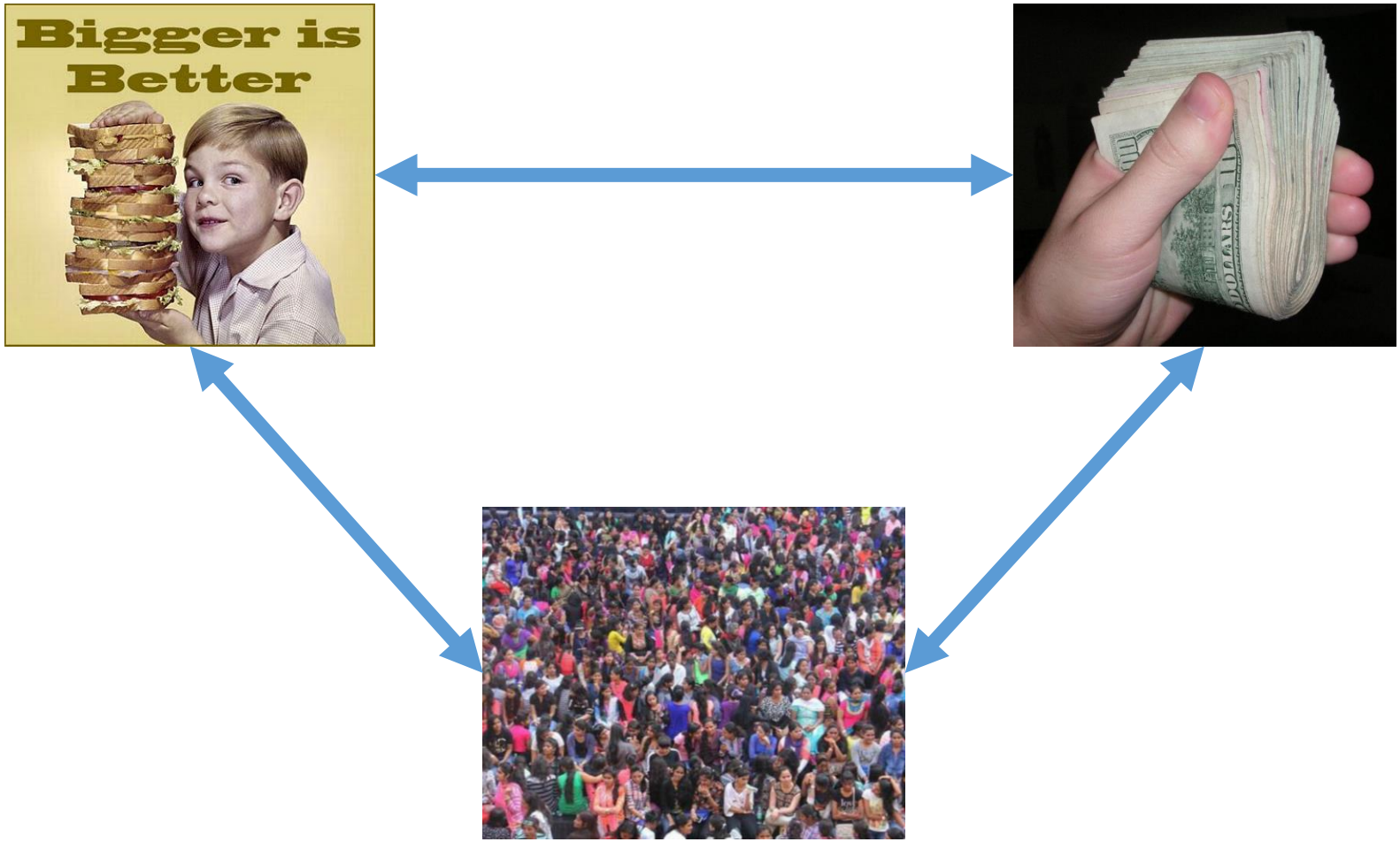
In pairs of two or groups of three:

- Add **three own examples** from your own experience (professional or other context)
- Chose some of the examples and discuss for each one:
  - Is this a **good sample**? Why?
  - If we want to make it a good sample, what would be important to watch out for?

15-20 minutes

Debrief: What makes a good sample?

What makes a good sample?



What makes a good sample?

- Size: the bigger the better?

What makes a good sample?

- Price: The cheaper the better?

What makes a good sample?

Representation of the population

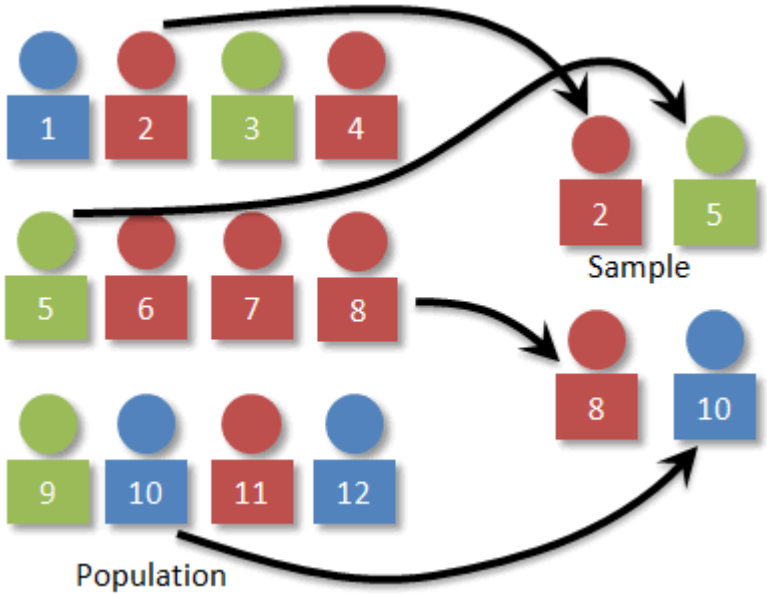
- In all possible regards?
- In all regards relevant to the questions of interest!
- But what are the questions of interest?
  → role of theory/previous knowledge

Sampling problems to watch out for

- Voluntary Response Samples
  - Sample has choice to respond to survey or not
- Convenience Samples
  - Sample chosen based on convenience
- Biased Samples
  - Sampling distorts population proportions in meaningful way
- Undercoverage
  - Chosen sample too small for selected design/analysis
- No response
  - response not high enough

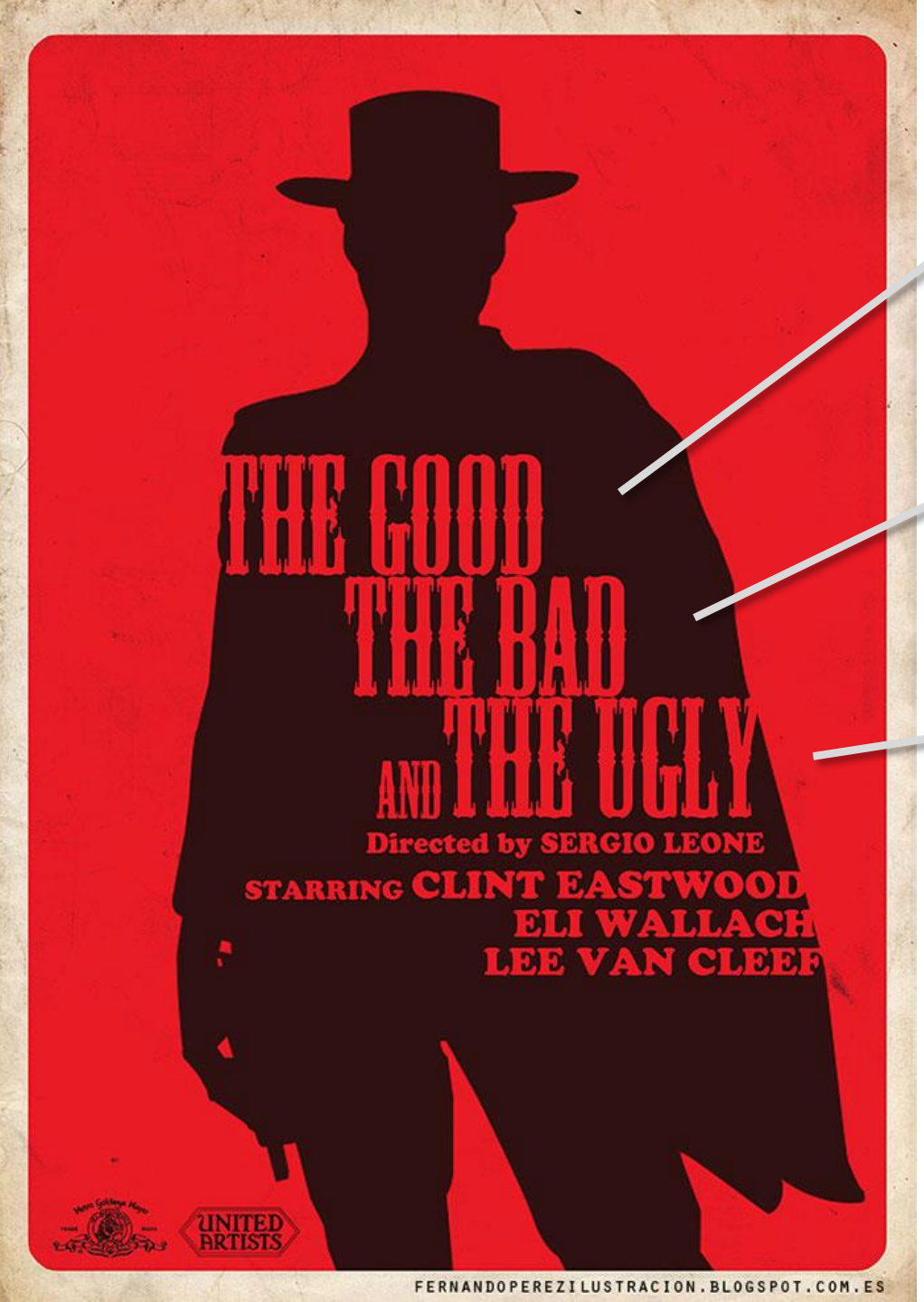
Counter measures

Random sampling



Stratified sampling

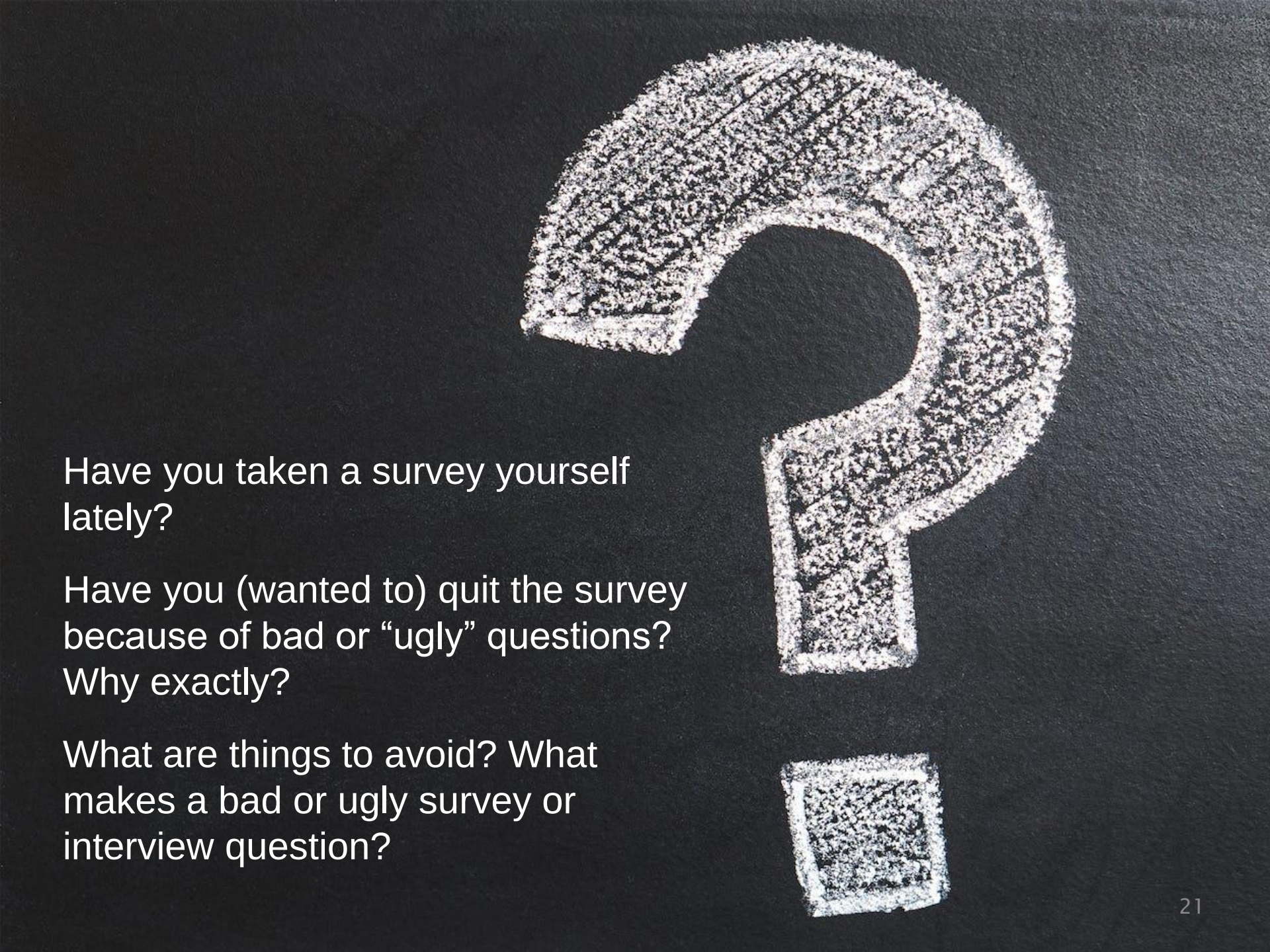
Surveys:
the good, the bad, and the ugly



HOW TO DO IT

HOW NOT TO DO IT

OK, BUT…



Have you taken a survey yourself lately?

Have you (wanted to) quit the survey because of bad or “ugly” questions? Why exactly?

What are things to avoid? What makes a bad or ugly survey or interview question?

What makes a good survey question?

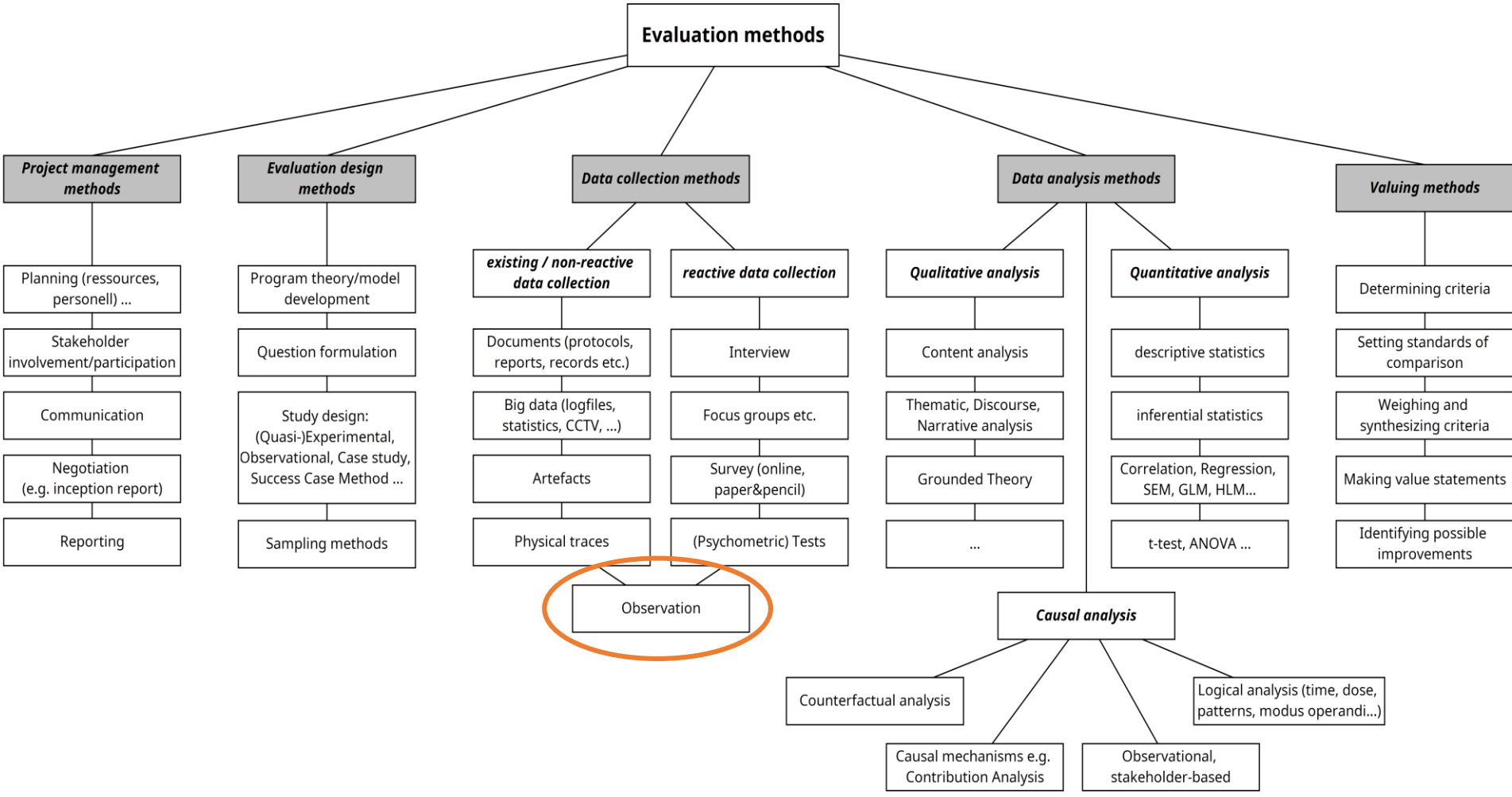
In pairs of two or groups of three

- Survey questions group exercise (2 pages)
- Checklist (1 page)

20–30 minutes

Debrief: What was unclear or seemed debatable?

The road less traveled: Seldom used methods


Evaluation methods
Project management methods
Planning (ressources, personell) ...
Stakeholder involvement/participation
Communication
Negotiation (e.g. inception report)
Reporting
Evaluation design methods
Program theory/model development
Question formulation
Study design: (Quasi-)Experimental, Observational, Case study, Success Case Method ...
Sampling methods
Data collection methods
existing / non-reactive data collection
Documents (protocols, reports, records etc.)
Big data (logfiles, statistics, CCTV, ...)
Artefacts
Physical traces
reactive data collection
Interview
Focus groups etc.
Survey (online, paper&pencil)
(Psychometric) Tests
Observation
Data analysis methods
Qualitative analysis
Content analysis
Thematic, Discourse, Narrative analysis
Grounded Theory
...
Quantitative analysis
descriptive statistics
inferential statistics
Correlation, Regression, SEM, GLM, HLM...
t-test, ANOVA ...
Causal analysis
Counterfactual analysis
Causal mechanisms e.g. Contribution Analysis
Observational, stakeholder-based
Logical analysis (time, dose, patterns, modus operandi...)
Valuing methods
Determining criteria
Setting standards of comparison
Weighing and synthesizing criteria
Making value statements
Identifying possible improvements

Observation for data collection

- Direct (in situ) vs. indirect (recording)
- Open vs. covert
- Participating vs. non participating
- Structured vs. unstructured
- Event vs. interval sampling

Interval Recording Form

Student's Name _____

Target Behavior: Rodney taps his feet, flaps his hands, or engages in other self stimulatory behavior.

Observation Length: 1 hour **Interval Length: 5 minutes**

date	begin time																				total %

date	begin time																				total %

Observation Data Collection

WILLIAM JESSUP UNIVERSITY

Candidate _____

Co-op Master Teacher _____

Date _____ Duration _____

School _____

Subject Area(s)		Type of Class	Type of School	Grade	
☐ Arts	☐ Science	☐ Mainstream	☐ Charter	☐ K-2	☐ 9-12
☐ Language Arts/ Reading	☐ History/ Social Science	☐ English Language Learners	☐ Private	☐ 3-5	
☐ Mathematics	☐ P.E/Health	☐ Special Education	☐ Public	☐ 6-8	

Student Information			
Number of ELL Students		Number of GATE Students	
Number of SN Students		Total Students in Class	

Character Traits Observed		
☐ Love	☐ Patience	☐ Faithfulness
☐ Joy	☐ Kindness	☐ Gentleness
☐ Peace	☐ Goodness	☐ Self-Control

Identify standard(s) addressed

(CA Content Standards & Frameworks: http://www.cde.ca.gov/be/st/ss/)

List materials used and identify if California State Board (SBE) adopted materials

(SBE Adopted materials: http://www3.cde.ca.gov/impricelist/implsearch.aspx)

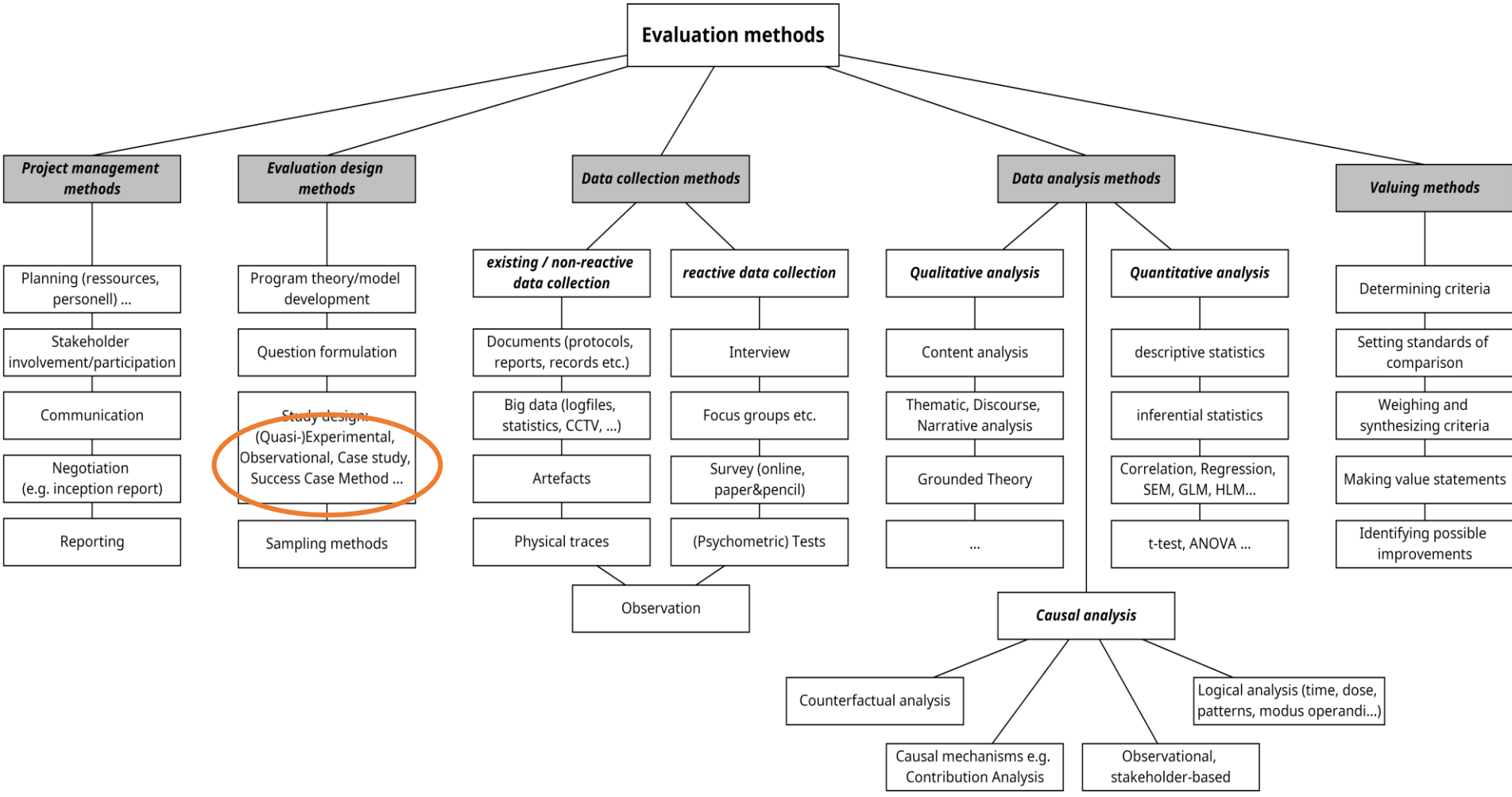
Observation for data collection

Pros

- Real world access
- Objective measurement of behavior

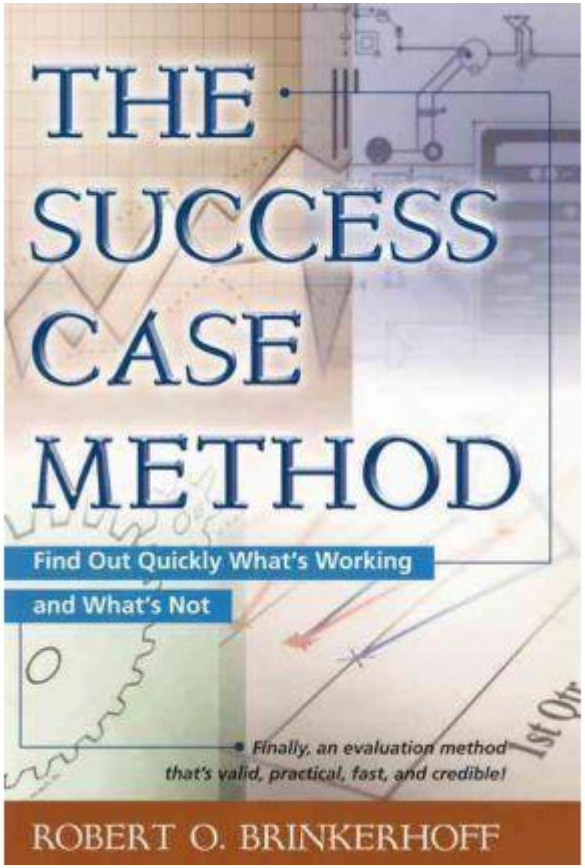
Cons

- Time consuming
- Observer bias, can be dependent on interpretation
- Only for certain kinds of criteria


Evaluation methods
Project management methods
Planning (ressources, personell) ...
Stakeholder involvement/participation
Communication
Negotiation (e.g. inception report)
Reporting
Evaluation design methods
Program theory/model development
Question formulation
Study design: (Quasi-)Experimental, Observational, Case study, Success Case Method ...
Sampling methods
Data collection methods
existing / non-reactive data collection
Documents (protocols, reports, records etc.)
Big data (logfiles, statistics, CCTV, ...)
Artefacts
Physical traces
reactive data collection
Interview
Focus groups etc.
Survey (online, paper&pencil)
(Psychometric) Tests
Observation
Data analysis methods
Qualitative analysis
Content analysis
Thematic, Discourse, Narrative analysis
Grounded Theory
...
Quantitative analysis
descriptive statistics
inferential statistics
Correlation, Regression, SEM, GLM, HLM...
t-test, ANOVA ...
Causal analysis
Counterfactual analysis
Causal mechanisms e.g. Contribution Analysis
Observational, stakeholder-based
Logical analysis (time, dose, patterns, modus operandi...)
Valuing methods
Determining criteria
Setting standards of comparison
Weighing and synthesizing criteria
Making value statements
Identifying possible improvements

Success case method

Look for where it worked
and where it did not work
and learn from the contrast



Success case method

Two phases:

1. Find successful and unsuccessful cases
   Case = target group member
   (people, institutions, regions etc.)

2. Conduct in-depth interviews

Success case method in detail

1. Focus and plan the study
2. Build an impact model

   → What makes a „success“ case?
3. Conduct broad study to find success cases

   Identify „best“ and „worst“ cases indirectly or directly
4. Interviews and documentation

   Find reasons for (non)success
5. Report and recommendations

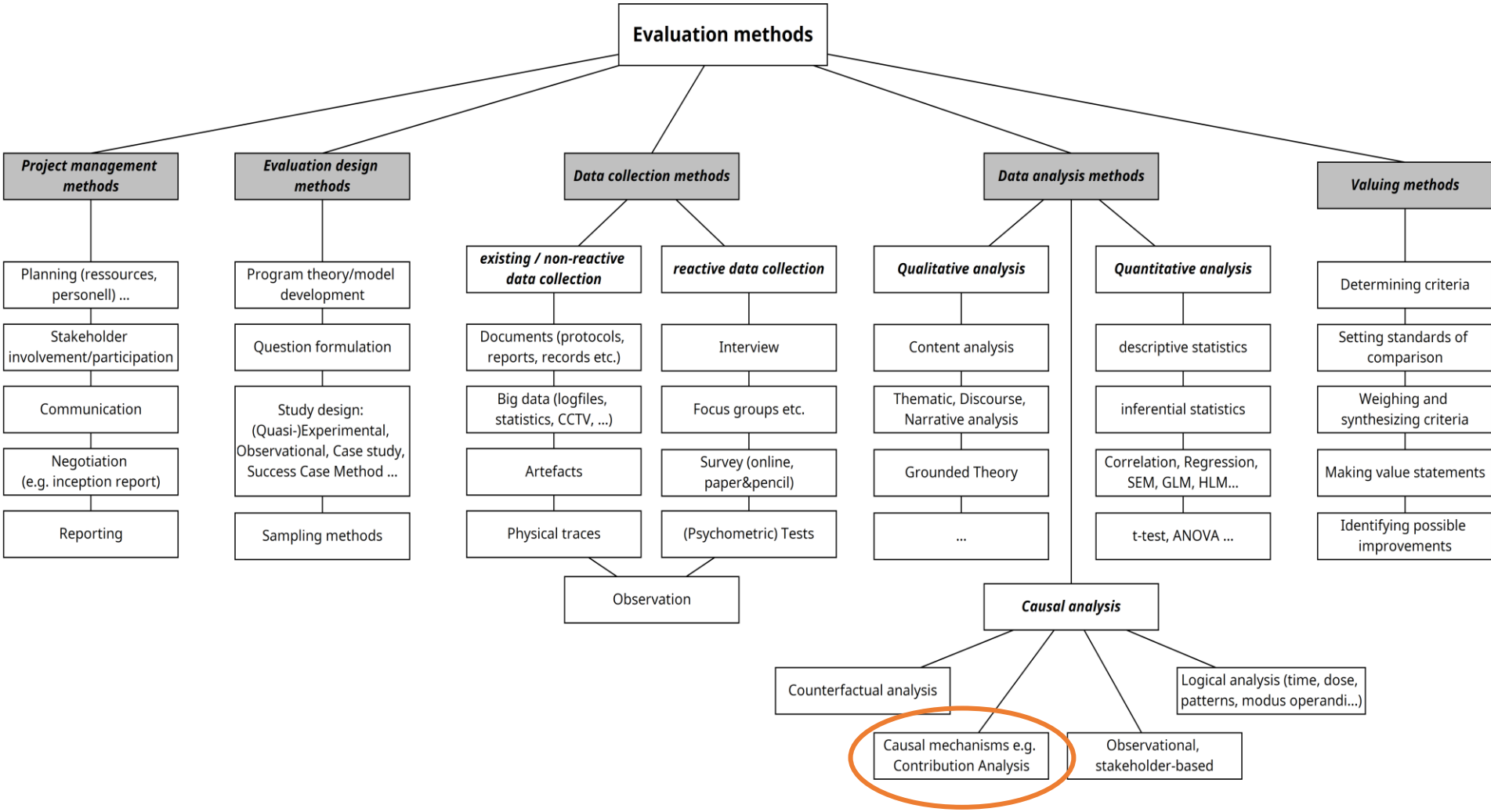
Success case method

Pros

- Pragmatic approach
- Efficient
- Aimed at learning about success and non-success

Cons

- Dependent on existence of extreme cases
- Mainly for formative, less for summative evaluation questions

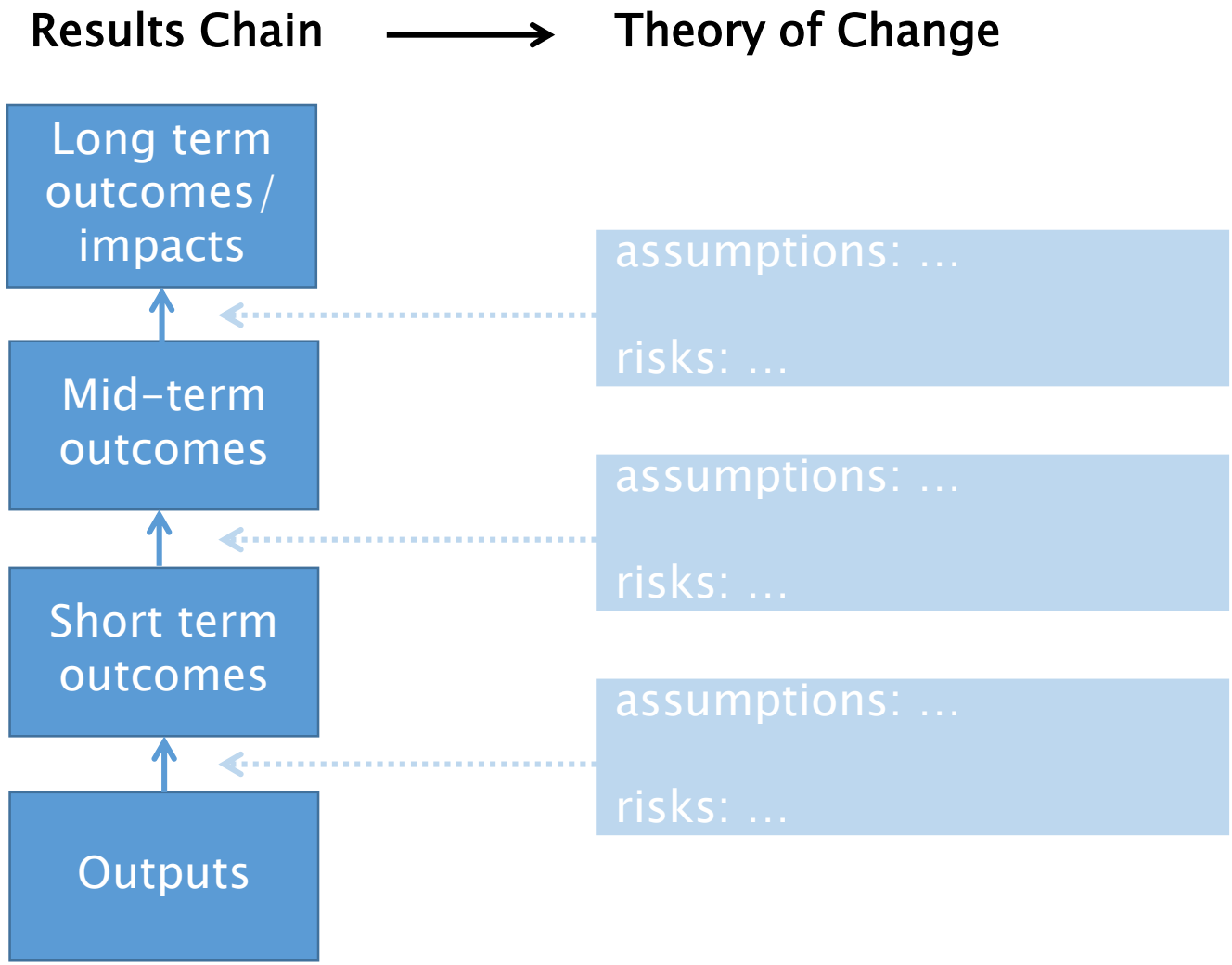

Evaluation methods
Project management methods
Planning (ressources, personell) ...
Stakeholder involvement/participation
Communication
Negotiation (e.g. inception report)
Reporting
Evaluation design methods
Program theory/model development
Question formulation
Study design: (Quasi-)Experimental, Observational, Case study, Success Case Method ...
Sampling methods
Data collection methods
existing / non-reactive data collection
Documents (protocols, reports, records etc.)
Big data (logfiles, statistics, CCTV, ...)
Artefacts
Physical traces
reactive data collection
Interview
Focus groups etc.
Survey (online, paper&pencil)
(Psychometric) Tests
Observation
Data analysis methods
Qualitative analysis
Content analysis
Thematic, Discourse, Narrative analysis
Grounded Theory
...
Quantitative analysis
descriptive statistics
inferential statistics
Correlation, Regression, SEM, GLM, HLM...
t-test, ANOVA ...
Causal analysis
Counterfactual analysis
Causal mechanisms e.g. Contribution Analysis
Observational, stakeholder-based
Logical analysis (time, dose, patterns, modus operandi...)
Valuing methods
Determining criteria
Setting standards of comparison
Weighing and synthesizing criteria
Making value statements
Identifying possible improvements

Contribution analysis (Mayne, 2008, 2012)

- Can we attribute observed effects of a program to the program?

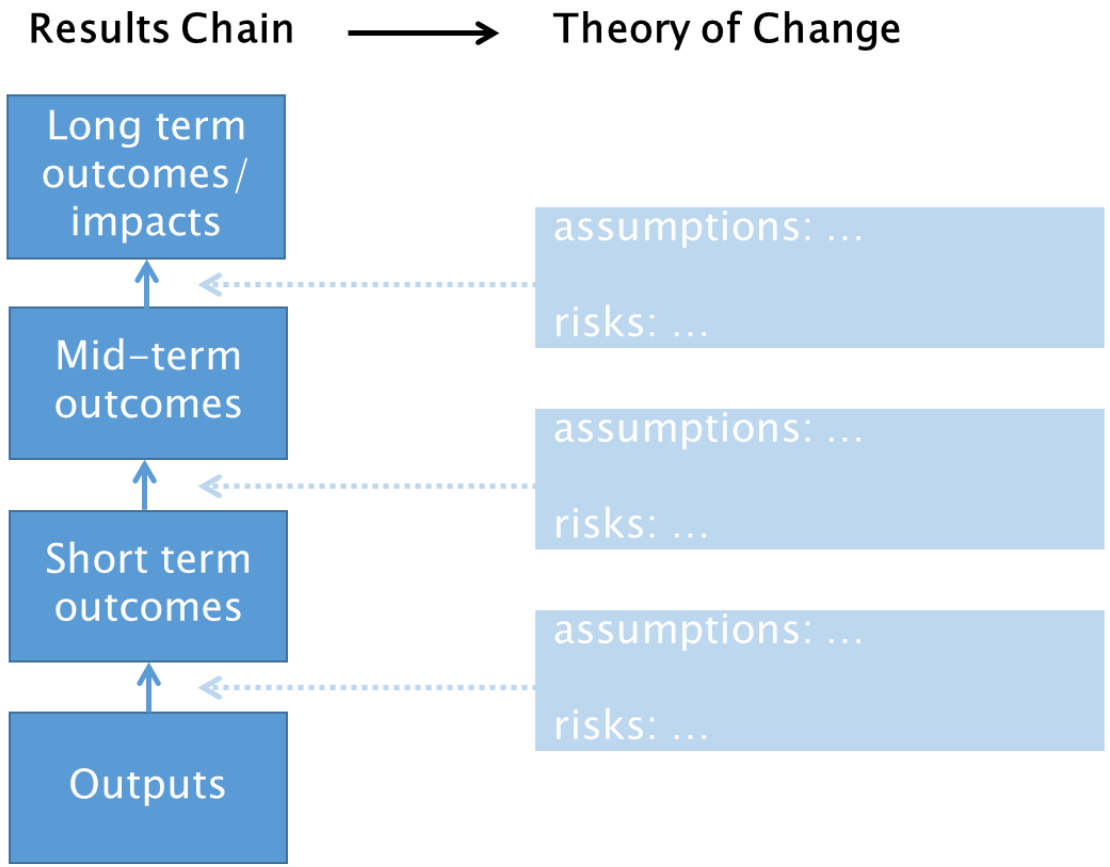
- Usual approach: (Quasi-)experimental designs („gold standard“)

Contribution analysis



Steps in contribution analysis

1. Identify problem
2. Develop „Theory of Change“



Steps in contribution analysis

3. Review existing evidence
   - What supports the ToC?
   - What are possible alternative explanations?

4. Develop initial „contribution story“
   - Why is it appropriate to assume that the program will contribute to intended effects?
   - How good is the existing evidence?
   - Do stakeholders agree?
   - What are the weak points?

Steps in contribution analysis

5. Collect additional evidence
   - Conventional data collection
   - Focus on weak points of the contribution story

6. Revise contribution story
   - Incorporate empirical evidence

Contribution analysis

Pros

- Alternative to experimental designs
- High information density
- Provokes theory development
- Potentially more efficient data collection

Cons

- Effort for literature review
- Effort for theory of change development
- Credibility?

How to mix the perfect Martini?
Measuring the right things right

Getting the right mix

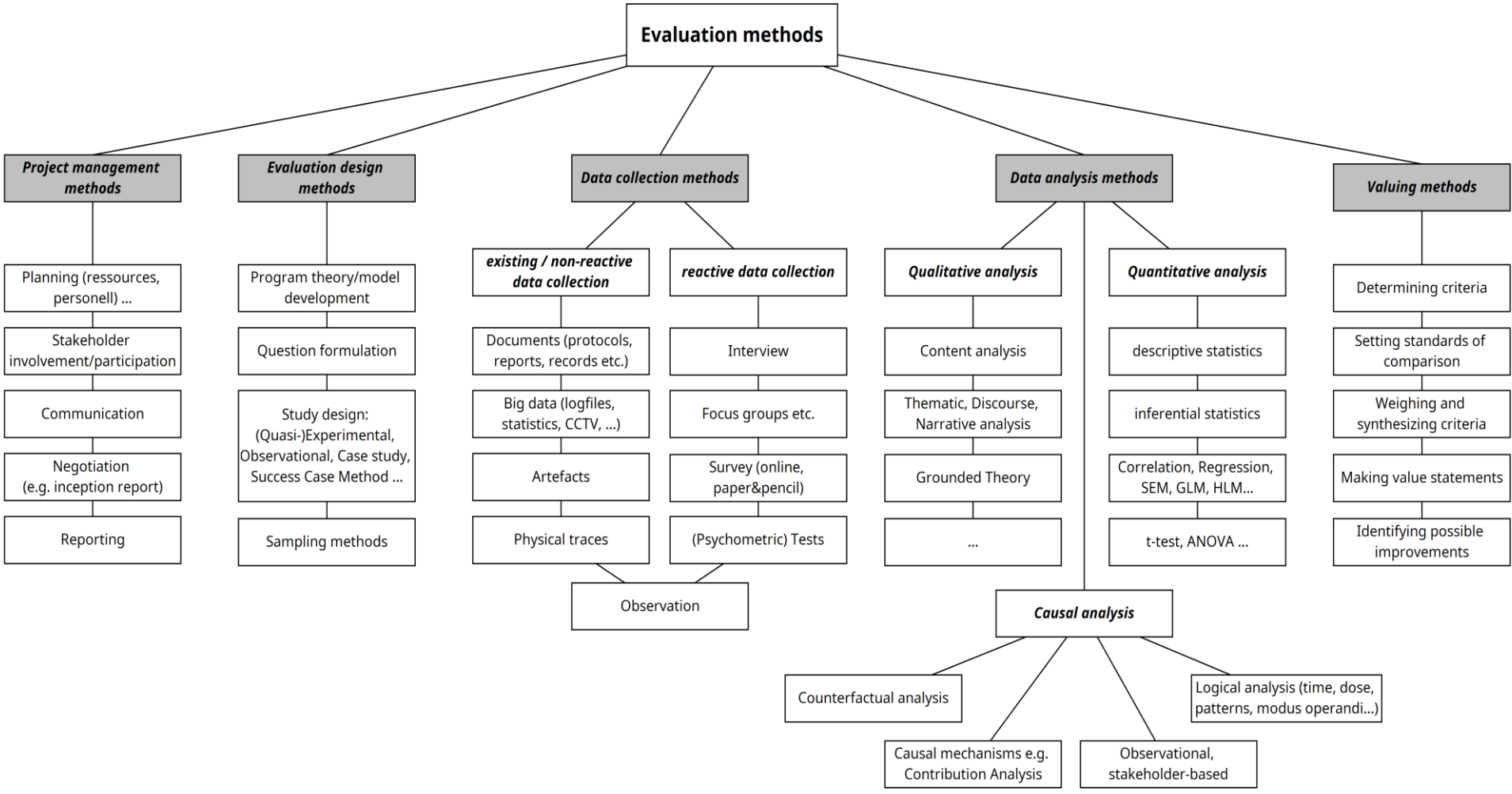
In groups of 4 to 6:

- Chose a specific **example study** from a volunteer group member
  - Preferably chose a study in its very **early phase**, with no decisions yet on evaluation methods
- Sketch the main elements of the study
- Discuss:
  - What data collection methods will be applicable in this setting? (for each of the indicators
  - What other evaluation methods seem appropriate?

20 minutes

Debrief: What directs our choice of methods?

The right mix?



„Mixed methods“?

- Mixed methods ≠ mixing methods
- How do methods complement each other?
- What if findings from different methods contradict?
- Additive vs. integrated information

Things to consider

- What level of credibility is needed?

„Gold Standard“
Randomized Controlled Studies (RCTs)

Randomized controlled trials (RCTs)

„Gold Standard“ of impact research

	t1	CM	t2
Trial	T_{t1}	X	T_{t2}
Control	C_{t1}	-	C_{t2}

Causal interpretation possible, due to

1. Covariation of suspected cause and effect
2. Cause before effect
3. Other causal influences ruled out by randomization

Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials

Gordon C S Smith, Jill P Pell

BMJ VOLUME 327 20–27 DECEMBER 2003 bmj.com

Parachutes reduce the risk of injury after gravitational challenge, but their effectiveness has not been proved with randomised controlled trials

What is already known about this topic

Parachutes are widely used to prevent death and major injury after gravitational challenge

---

Parachute use is associated with adverse effects due to failure of the intervention and iatrogenic injury

---

Studies of free fall do not show 100% mortality

What this study adds

No randomised controlled trials of parachute use have been undertaken

---

The basis for parachute use is purely observational,

"Those who advocate evidence based medicine and criticise use of interventions that lack an evidence base will not hesitate to demonstrate their commitment by **volunteering for a double blind, randomised, placebo controlled, crossover trial**."

Things to consider

- What level of credibility is needed?

- Do we rely on only one source of data?
  - Multiple data collection methods
  - Multiple perspectives

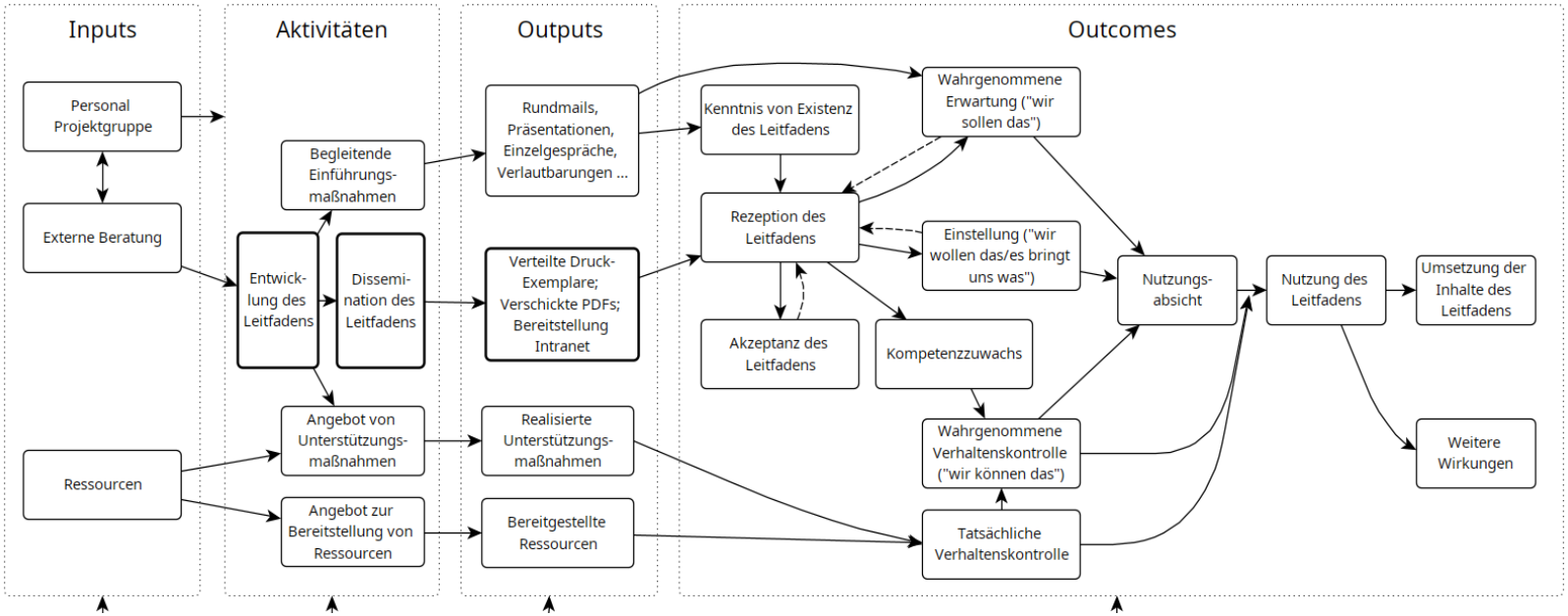
  →Triangulation

- Are we asking the right questions?

  → Role of program theory

Role of program theory

Ressourcen	Aktivitäten	Outputs	Kurzfristige Outcomes	Längerfristige Outcomes	Impacts
• Qualifizierte Trainerinnen und Trainer • Räumlichkeiten • Schulungsmaterial	• Weiterbildung für pädagogische Fachkräfte	• Anzahl an durchgeführten Weiterbildungen • Teilnahmestunden	• Akzeptanz der Weiterbildung • Wissenszuwachs • Einstellungsänderung • Umsetzungs-motivation	• Umsetzung der Weiterbildungsinhalte in der päd. Praxis mit Kindern • Weitergabe des Gelernten innerhalb der päd. Einrichtung • Weiterempfehlung des Angebots	• Verbesserter Bildungserfolg der Kinder • Auswirkungen auf die päd. Einrichtung • Aufwertung des gesellschaftlichen Stellenwerts des Weiterbildungsthemas



Things to consider

- What level of credibility is needed?
- Do we rely on only one source of data?
  - Multiple data collection methods
  - Multiple perspectives

  →Triangulation
- Are we asking the right questions?

  → Role of program theory
- Do we measure the right things right?

  → Question of criteria

Selecting evaluation criteria

What features of a program make it a success?

- Attention of goals (intended outcomes)?
  - Then what about unintended consequences?
- ROI: bang for the buck?
- Generic criteria?
  - e.g. OECD DAC:
    relevance, coherence, effectiveness, efficiency, sustainability

Final words:
Formative vs. summative evaluation

- Measuring success vs. understanding successs

- Often it‘s not realistic to expect innovations to work from the start.

  → Don‘t waste all your ressources on finding out what worked.

  → Try to also understand why it work and how it can be made to work in different circumstances in the future.

Review of the course

- Evaluation methods: the big picture
- Sampling: problems and strategies
- Survey questions: the good, the bad, and the ugly
- Less common methods:
    1. Observational methods
    2. Case studies / Success case method
    3. Contribution analysis
- Measuring the right things right


Thank you!
mail@jan-hense.de