SURVEY RESEARCH METHODS IN SOFTWARE ENGINEERING

Prof. Marcos Kalinowski

kalinowski@inf.puc-rio.br
http://www.inf.puc-rio.br/~kalinowski
@prof_kalinowski

DEPARTAMENTO DE INFORMÁTICA
PUC-RIO

Introductions ...

- Professor at the Department of Informatics (DI) of PUC-Rio
- Coordinator of the ExACTa PUC-Rio Lab @ DI PUC-Rio
- M.Sc And Ph.D level advisor in the areas of Software Engineering and Data Science
- Member of the ISERN *(International Software Engineering Research Network)*

kalinowski@inf.puc-rio.br
http://www.inf.puc-rio.br/~kalinowski
@prof_kalinowski
http://www.linkedin.com/in/kalinowski/

Among the top-ranked universities and **leader in industry integration in Latin America**, according to Times Higher Eduction

First Brazilian graduate program in Computing, assessed with the **highest research quality score (7/7)** by the Brazilian Federal Government

R&D in **Software Engineering**, **Data Science**, and **Human-Computer Interaction**

Agenda

1

INTRODUCTION TO SURVEY RESEARCH

- Characteristics and Purposes
- Examples of Surveys

2

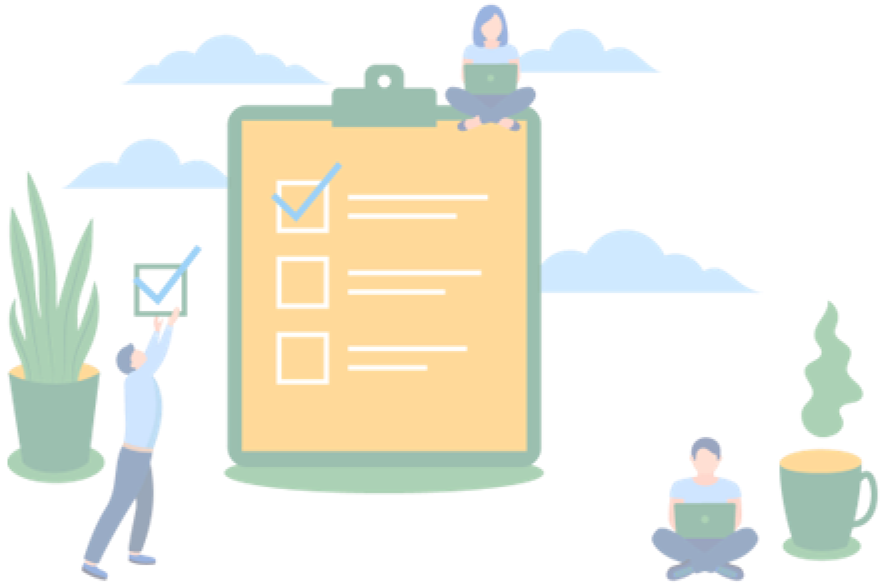
SURVEY DESIGN

- Basics of Survey Design
- Goal-Question-Metric Driven Design
- Theory-Driven Design
- Issues When Assessing Psychological Constructs
- Survey Instrument Evaluation

3

SAMPLING

- Representativeness
- Sample Size Estimation
- Ethics



Agenda

4 DATA COLLECTION
- Invitation and Follow-Up

5 STATISTICAL ANALYSIS
- Descriptive Statistics
- Null-hypothesis Significance Testing
- Bootstrapping Confidence Intervals
- Bayesian Analysis
- Structural Equation Modeling

6 QUALITATIVE ANALYSIS
- Grounded Theory Procedures

7 THREATS TO VALIDITY AND RELIABILITY

8 HANDS-ON EXERCISE

Bibliography

Challenges in Survey Research

Stefan Wagner, Daniel Mendez, Michael Felderer, Daniel Graziotin, and Marcos Kalinowski

Wagner, S., Mendez, D., Felderer, M., Graziotin, D. and Kalinowski, M., 2020. **Challenges in survey research.** In: Contemporary Empirical Methods in Software Engineering (pp. 93-125). Springer, Cham.

Fowler Jr, F.J., 2013. **Survey research methods.** Sage publications.

Linåker, J., Sulaman, S.M., Maiani de Mello, R. and Höst, M., 2015. **Guidelines for conducting surveys in software engineering**. Technical Report.

Kitchenham, B.A. and Pfleeger, S.L., 2008. **Personal opinion surveys.** In Guide to advanced empirical software engineering (pp. 63-92). Springer, London.

+ Specific methodological and survey articles cited throughout the course.

Acknowledgments


Challenges in Survey Research

Stefan Wagner, Daniel Mendez, Michael Felderer, Daniel Graziotin, and Marcos Kalinowski


S. Wagner


D. Mendez


M. Felderer


D. Graziotin


T. Baldassarre


S. Baltes


S. Biffl


T. Conte


M. Genero


M. Kuhrmann


E. Mendes


R. Prikladnicki


P. Ralph


M.A. Storey


G. Travassos


C. Wohlin

INTRODUCTION TO SURVEY RESEARCH

Characteristics

- Surveys are probably the most commonly used research method worldwide

- Surveys are conducted when a phenomena (*e.g.*, the use of a technique or tool) **already has taken place** or **before it occurs**
  - A *survey* provides **no control of the execution or measurement**
    - I.e., it is not possible to manipulate variables as in the other investigation methods
  - Surveys should aim at obtaining the largest amount of understanding from the fewest number of variables since this reduction also eases the data collection and analysis

- Surveys are almost never conducted to create an understanding concerning a particular **sample**, the typical focus is on generalizing results to the **population** from which the sample was drawn.

Purpose

- General objectives for conducting a survey *(Wohlin et al., 2012; Wagner et al., 2020)*:

EXPLORATIVE SURVEYS	DESCRIPTIVE SURVEYS	EXPLANATORY SURVEYS
are used as a pre-study to a more thorough investigation to assure that important issues are not forgotten (e.g., <u>constructs</u> in a theory like requirements elicitation techniques)	can be conducted to enable assertions about some population like the distribution of certain attributes (e.g., usage of requirements elicitation techniques)	aim at making explanatory claims about the population (e.g., why specific requirements elicitation techniques are used in specific contexts)

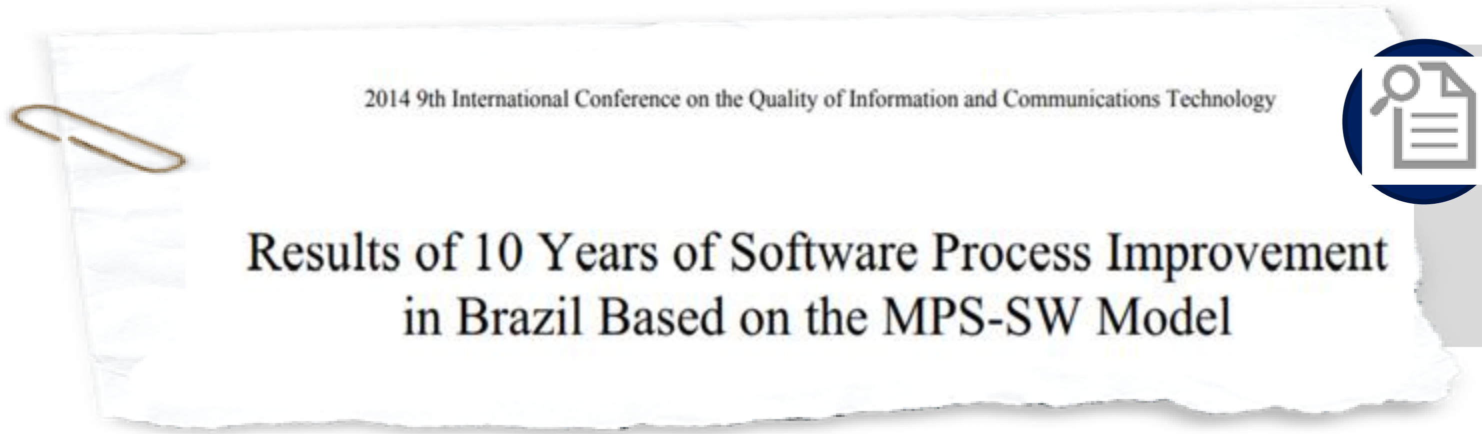
Characteristics and Purpose

"Theory building and evaluation can guide the design and analysis of surveys, and surveys can also be applied to test theories."

(Wagner et al., 2020)

Wagner, S., Mendez, D., Felderer, M., Graziotin, D. and Kalinowski, M., 2020. **Challenges in survey research.** In: Contemporary Empirical Methods in Software Engineering (pp. 93-125). Springer, Cham.

Examples of Surveys


2014 9th International Conference on the Quality of Information and Communications Technology

Results of 10 Years of Software Process Improvement in Brazil Based on the MPS-SW Model

Kalinowski, M., Weber, K., Franco, N., Barroso, E., Duarte, V., Zanetti, D. and Santos, G., 2014, September. **Results of 10 years of software process improvement in Brazil based on the MPS-SW model.** In 2014 9th International Conference on the Quality of Information and Communications Technology (pp. 28-37).

Kalinowski, M., Weber, K., Santos, G., Franco, N., Duarte, V. and Travassos, G., 2015. **Software Process Improvement Results in Brazil Based on the MPS-SW Model.** Software Quality Professional, 17(4): 15-28.

Travassos, G.H. and Kalinowski, M., 2014. **iMPS 2013: Evidence on performance of organizations that adopted the MPS-SW.** Campinas, Brazil: Softex.

https://softex.br/mpsbr/pesquisas-mps/

Examples of Surveys


Empir Software Eng (2017) 22:2298–2338
DOI 10.1007/s10664-016-9451-7

CrossMark

Naming the pain in requirements engineering

Contemporary problems, causes, and effects in practice

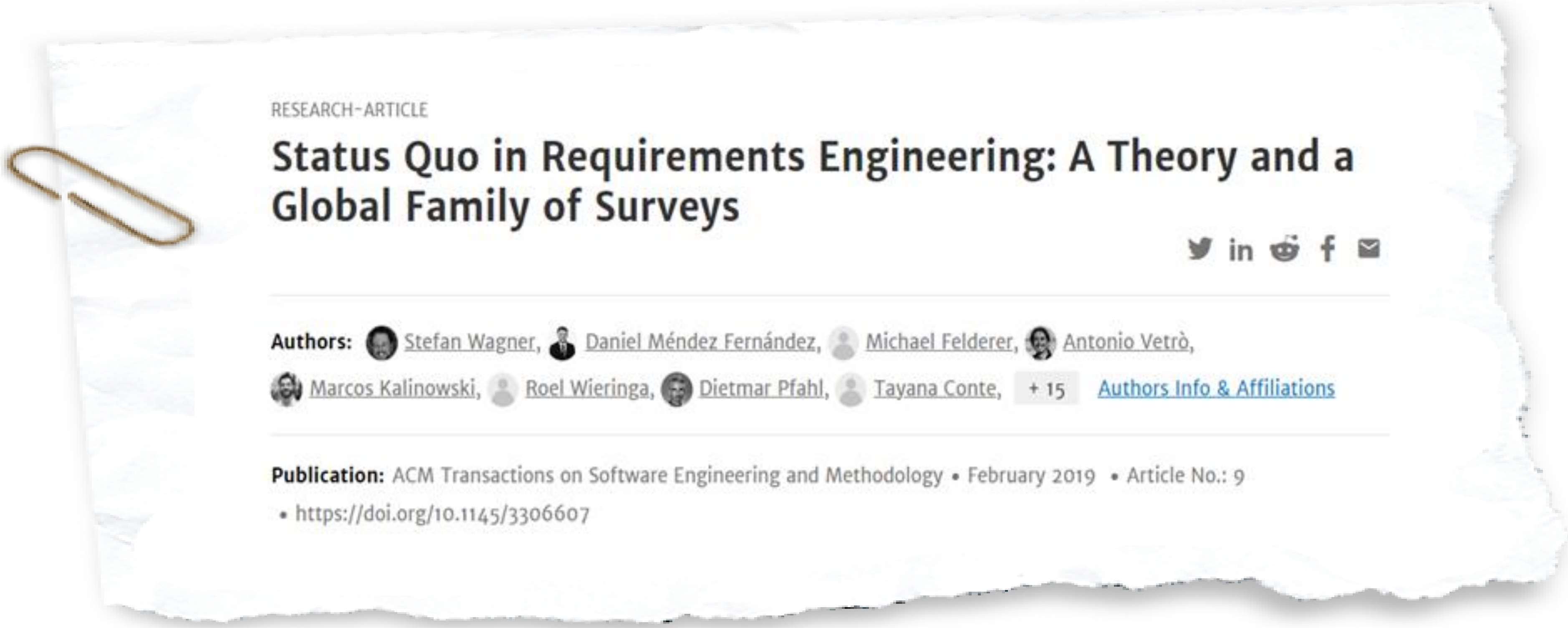
D. Méndez Fernández[1] · S. Wagner[2] · M. Kalinowski[3] · M. Felderer[4] · P. Mafra[3] · A. Vetrò[5] · T. Conte[6] · M.-T. Christiansson[7] · D. Greer[8] · C. Lassenius[9] · T. Männistö[10] · M. Nayabi[11] · M. Oivo[12] · B. Penzenstadler[13] · D. Pfahl[14] · R. Prikladnicki[15] · G. Ruhe[11] · A. Schekelmann[16] · S. Sen[17] · R. Spinola[18,19] · A. Tuzcu[20] · J. L. de la Vara[21] · R. Wieringa[22]

Published online: 24 October 2016
© Springer Science+Business Media New York 2016

Naming the Pain in Requirements Engineering
NaPiRE

Fernández, D. M.; Wagner, S.; Kalinowski, M.; Felderer, M.; Mafra, P.; Vetro, A.; Conte, T.; Christiansson, M.; Greer, D.; Lassenius, C.; Männistö, T.; Nayabi, M.; Oivo, M.; Penzenstadler, B.; Pfahl, D.; Prikladnicki, R.; Ruhe, G.; Schekelmann, A.; Sen, S.; Spínola, R. O.; Tuzcu, A.; de la Vara, J. L.; and Wieringa, R. **Naming the pain in requirements engineering - Contemporary problems, causes, and effects in practice**. Empirical Software Engineering, 22(5): 2298-2338. 2017.

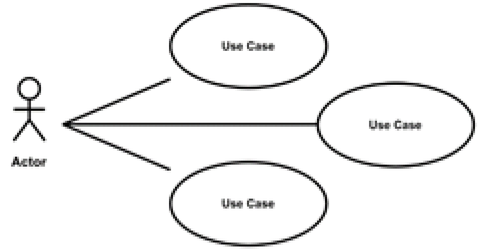
Examples of Surveys


RESEARCH-ARTICLE

Status Quo in Requirements Engineering: A Theory and a Global Family of Surveys

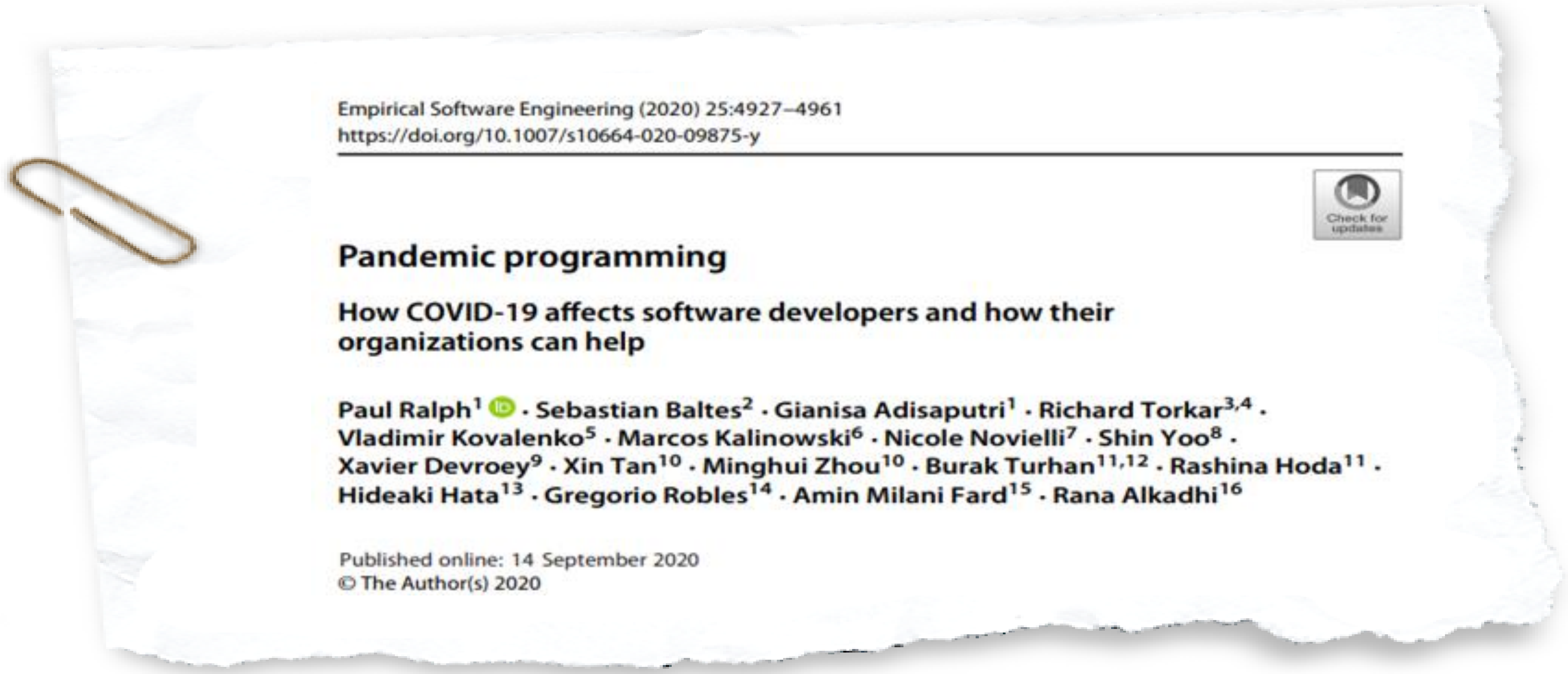
Authors: Stefan Wagner, Daniel Méndez Fernández, Michael Felderer, Antonio Vetrò, Marcos Kalinowski, Roel Wieringa, Dietmar Pfahl, Tayana Conte, + 15 Authors Info & Affiliations

Publication: ACM Transactions on Software Engineering and Methodology • February 2019 • Article No.: 9 • https://doi.org/10.1145/3306607



Wagner, S., Fernández, D. M., Felderer, M., Vetro, A., Kalinowski, M., Wieringa, R., Pfahl, D., Conte, T., Christiansson, M., Greer, D., Lassenius, C., Männistö, T., Nayebi, M., Oivo, M., Penzenstadler, B., Prikladnicki, R., Ruhe, G., Schekelmann, A., Sen, S., Spínola, R.O., Tuzcu, A., de la Vara, J. L., and Winkler, D, **Status Quo in Requirements Engineering: A Theory and a Global Family of Surveys.** ACM Transactions on Software Engineering and Methdology, 28(2): 9:1-9:48. 2019.

Examples of Surveys



Empirical Software Engineering (2020) 25:4927–4961
https://doi.org/10.1007/s10664-020-09875-y

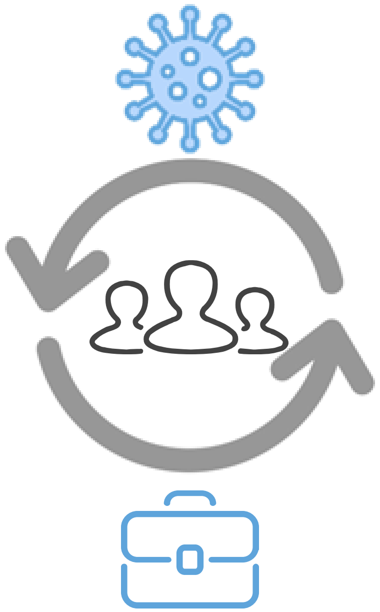
Check for updates

Pandemic programming

How COVID-19 affects software developers and how their organizations can help

Paul Ralph[1] · Sebastian Baltes[2] · Gianisa Adisaputri[1] · Richard Torkar[3,4] · Vladimir Kovalenko[5] · Marcos Kalinowski[6] · Nicole Novielli[7] · Shin Yoo[8] · Xavier Devroey[9] · Xin Tan[10] · Minghui Zhou[10] · Burak Turhan[11,12] · Rashina Hoda[11] · Hideaki Hata[13] · Gregorio Robles[14] · Amin Milani Fard[15] · Rana Alkadhi[16]

Published online: 14 September 2020
© The Author(s) 2020

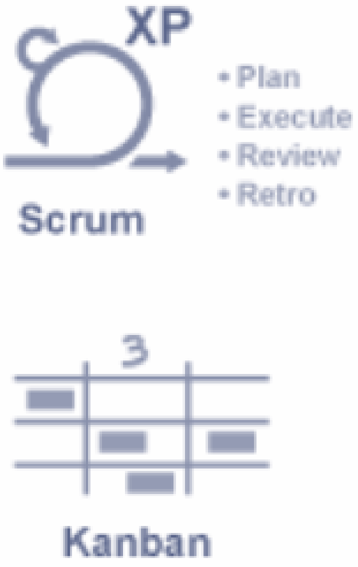


Ralph, P., Baltes, S., Adisaputri, G., Torkar, R., Kovalenko, V., Kalinowski, M., Novielli, N., Yoo, S., Devroey, X., Tan, X., Zhou, M., Turhan, B., Hoda, R., Hata, H., Robles, G., Fard, A. M., and Alkadhi, R, **Pandemic Programming How COVID-19 affects software developers and how their organizations can help**. Empirical Software Engineering (2020), 25: 4927-4961. 2020.

Examples of Surveys

What Makes Agile Software Development Agile?

Marco Kuhrmann, Paolo Tell, Regina Hebig, Jil Klünder, Jürgen Münch, Oliver Linssen, Dietmar Pfahl, Michael Felderer, Christian R. Prause, Stephen G. MacDonell, Joyce Nakatumba-Nabende, David Raffo, Sarah Beecham, Eray Tüzün, Gustavo López, Nicolas Paez, Diego Fontdevila, Sherlock A. Licorish, Steffen Küpper, Günther Ruhe, Eric Knauss, Özden Özcan-Top, Paul Clarke, Fergal McCaffery, Marcela Genero, Aurora Vizcaino, Mario Piattini, Marcos Kalinowski, Tayana Conte, Rafael Prikladnicki, Stephan Krusche, Ahmet Coşkunçay, Ezequiel Scott, Fabio Calefato, Svetlana Pimonova, Rolf-Helge Pfeiffer, Ulrik Pagh Schultz, Rogardt Heldal, Masud Fazal-Baqaie, Craig Anslow, Maleknaz Nayebi, Kurt Schneider, Stefan Sauer, Dietmar Winkler, Stefan Biffl, Maria Cecilia Bastarrica, and Ita Richardson



Kuhrmann, M., Tell, P., Hebig, R. *et al*. **What Makes Agile Software Development Agile?** Submitted to Transactions on Software Engineering (2021).

SURVEY DESIGN

Survey Design

- Basics of Survey Design
- Goal-Question-Metric-Driven Design
- Theory-Driven Design
- Issues When Assessing Psychological Constructs
- Survey Instrument Evaluation

QUESTIONNAIRE TYPES

- Self-administered questionnaire
- Interviewer-administered questionnaire

QUESTION CATEGORIES

- Demographic questions
- Substantive questions
- Filter questions
- Sensitive questions

QUESTION TYPES

- Open-ended
- Close-ended
- Hybrid questions

Measurement scales

More Information

Nominal
- Values can be counted

Ordinal
- Values can be counted and ordered

e.g. Likert

Interval
- Values can be counted and ordered
- Distance between values can be interpreted

Ratio
- Values can be counted and ordered
- Distance between values can be interpreted
- Radio between values can be interpreted

Conditions that must be fulfilled to get appropriate responses

Questions must be understandable by the target population

Respondents must have sufficient knowledge to answer

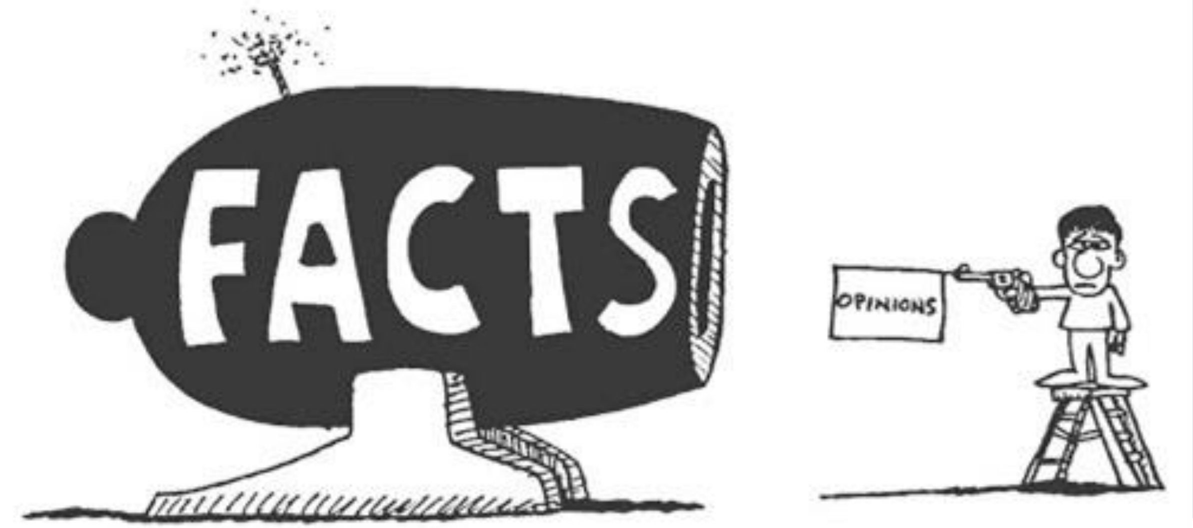
Participants must be motivated and willing to participate

Suggestions to avoid common question wording problems

(adapted from Kitchenham and Pfleeger, 2008)

- ✓ Using appropriate and simple language
- ✓ Avoiding technical terms
- ✓ Keeping questions short
- ✓ Avoiding vague sentences
- ✓ Avoiding sensitive questions
- ✓ Avoiding too demanding questions
- ✓ Avoiding double-barreled questions
- ✓ Avoiding double negatives
- ✓ Avoid asking about long gone events

In a survey, we can either ask for the **opinions** of the participants on topics or for specific **facts** that they experienced.

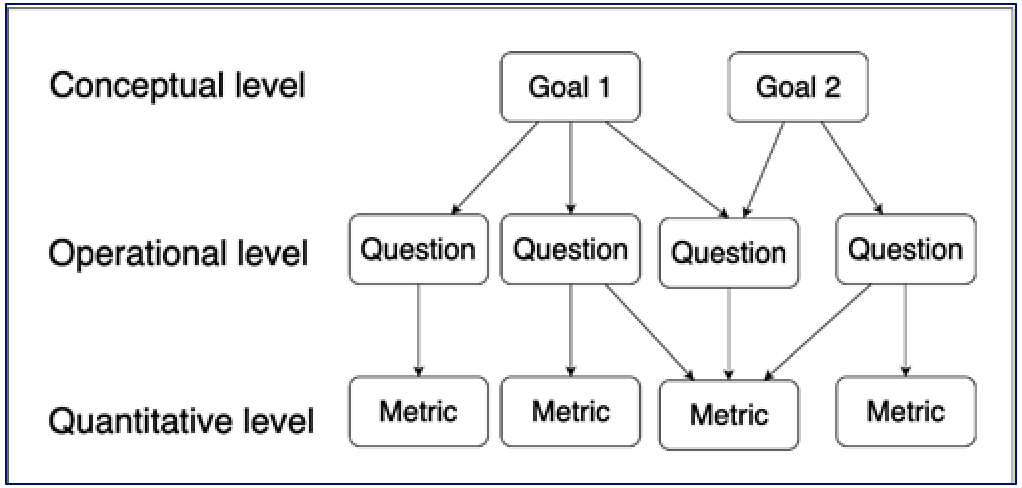


Goal-Question-Metric-Driven Design

- Based on the Goal Question Metric (GQM) Paradigm *(Basili and Romback, 1988)*

GQM defines a way to plan and execute measurement and analysis activities:

1. Starts with the declaration of the measurement, **Goals**
2. From the goals, **Questions** that we would like to answer with the data interpretation are defined
3. Finally, from the questions, the **Metrics** and the data to be collected are defined



Basili, V.R. and Rombach, H.D., 1988. **The TAME project: Towards improvement-oriented software environments.** IEEE Transactions on software engineering, 14(6), pp.758-773.

Goal Definition Template

Analyze <object of study>

with the purpose of <goal>

with respect to <quality focus>

from the point of view of the <perspective>

in the context of <context>

Measurement activities need clear goals
GQM: characterize, understand, evaluate, predict, improve.

Basili, V.R. and Rombach, H.D., 1988. **The TAME project: Towards improvement-oriented software environments.** IEEE Transactions on software engineering, 14(6), pp.758-773.


EXAMPLE

Goal Definition Template (Example)

GOAL

Analyze the profile of software development organizations

with the purpose of characterizing

with respect to the organizations' current profile, satisfaction degree regarding the MPS model, variation of presence in international markets, variation of exportation volume, and variation concerning cost, estimation accuracy, productivity, quality, user satisfaction, and return of investment (ROI)

from the point of view the software development organizations

in the context of software development organizations with unexpired MPS-SW assessments published in the SOFTEX portal

Kalinowski, M., Weber, K.C. and Travassos, G.H., 2008, October. **iMPS: an experimentation based investigation of a nationwide software development reference model.** In Proceedings of the Second ACM-IEEE international symposium on Empirical Software Engineering and Measurement (ESEM).

Further Goal-Question-Metric-Driven Design Examples

"Analyze Social BPM with the purpose of characterizing with respect to adoption of its practices and technologies during the BPM lifecycle from the point of view of BPM participants or managers In the context of Brazilian organizations."

Batista, M., Magdaleno, A. and Kalinowski, M., 2017, May. A Survey on the use of Social BPM in Practice in Brazilian Organizations. In Anais do XIII Simpósio Brasileiro de Sistemas de Informação (SBSI) (pp. 436-443). SBC.

"Analyze V&V methods with the purpose of characterization with respect to their suitability for addressing ISO 25010 software quality characteristics from the point of view of experts in the area of V&V in the context of the software engineering research community."

Mendoza, I., Kalinowski, M., Souza, U. and Felderer, M., 2019, January. Relating verification and validation methods to software product quality characteristics: results of an expert survey. In Proc. of the Software Quality Days Conference (SWQD) (pp. 33-44).

Goal Definition Template (Example)

GOAL

Analyze software development organizations

with the purpose of characterizing

with respect to the organizations' current profile, satisfaction degree regarding the MPS model, variation of presence in international markets, variation of exportation volume, and variation concerning cost, estimation accuracy, productivity, quality, user satisfaction, and return of investment (ROI)

from the point of view the software development organizations

in the context of software development organizations with unexpired MPS-SW assessments published in the SOFTEX portal

Kalinowski, M., Weber, K.C. and Travassos, G.H., 2008, October. **iMPS: an experimentation based investigation of a nationwide software development reference model.** In Proceedings of the Second ACM-IEEE international symposium on Empirical Software Engineering and Measurement (ESEM).

Question and Metrics (1/2)

QUESTION

Q1: What is the organization’s estimation accuracy?

METRICS

M1.1: Average Project Duration = Average duration of projects conducted within the last 12 months, measured in months.

M1.2: Average Project Estimated Duration = Average estimated duration of projects conducted within the last 12 months, measured in months.

M1.3: Estimation Accuracy = 1 - |((Average Project Duration – Average Project Estimated Duration) / Average Project Duration)|

Kalinowski, M., Weber, K.C. and Travassos, G.H., 2008, October. **iMPS: an experimentation based investigation of a nationwide software development reference model.** In Proceedings of the Second ACM-IEEE international symposium on Empirical Software Engineering and Measurement (ESEM).

Question and Metrics (2/2)

QUESTION

Q2: What is the organization's Return of Investment (RoI) of adopting MPS-SW?

METRICS

M2.1: Variation in net sales = Percentage of variation in net sales.

M2.2: Investment in implementing MPS = Percentage of net sales invested in implementing MPS

M2.3: Investment in assessing MPS = Percentage of net sales invested in the MPS assessment

M2.4: ROI = (Variation in net sales / (Investment in implementing MPS + Investment in assessing MPS)) * 100

Kalinowski, M., Weber, K.C. and Travassos, G.H., 2008, October. **iMPS: an experimentation based investigation of a nationwide software development reference model.** In Proceedings of the Second ACM-IEEE international symposium on Empirical Software Engineering and Measurement (ESEM).

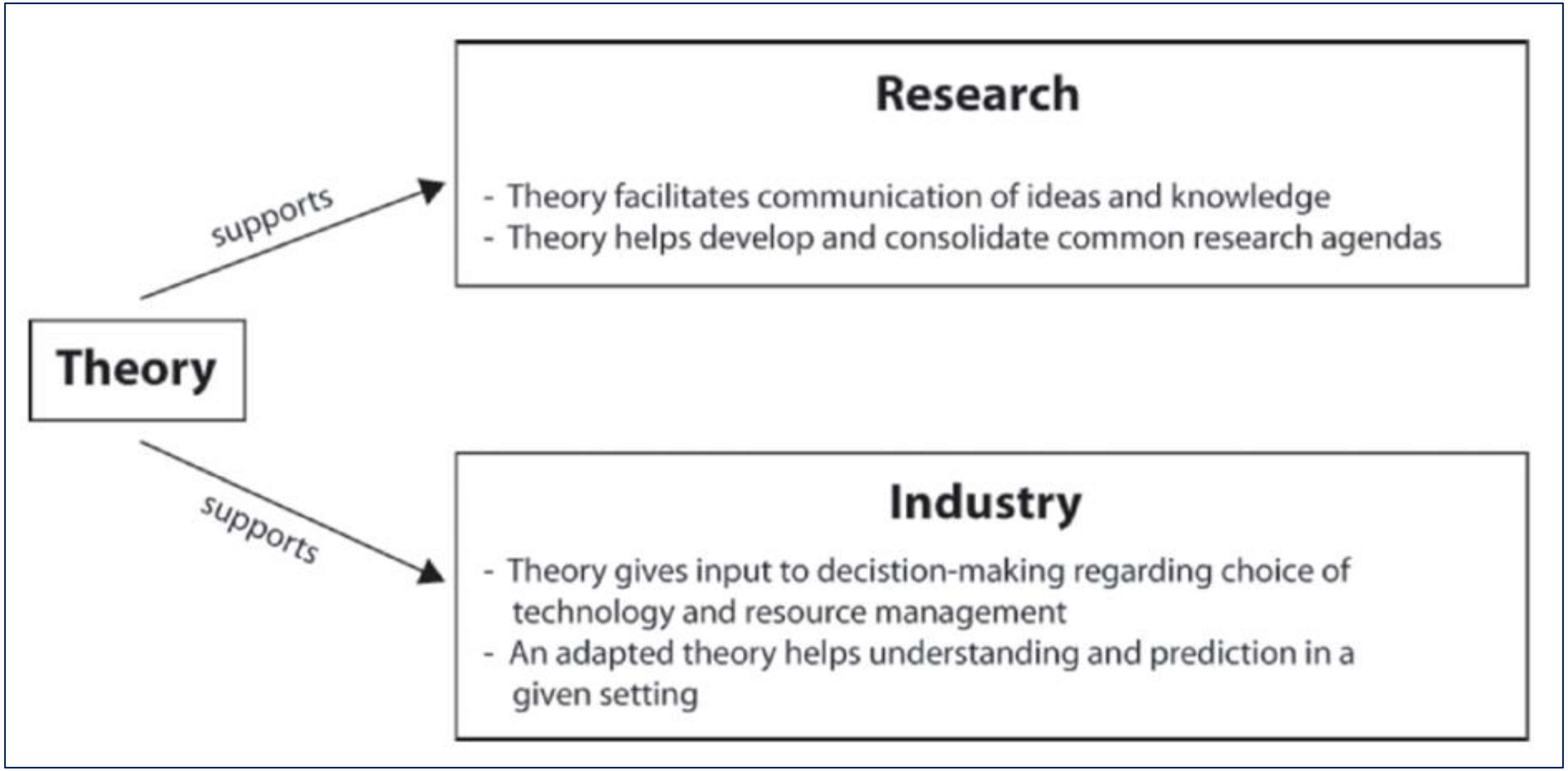
A theory provides explanations and understanding in terms of basic constructs and underlying mechanisms, which constitute an important counterpart to knowledge of passing trends and their manifestation *(Hannay et al. 2007)*:

- From the practical perspective, theories should be useful and explain or predict phenomena that occur in software engineering
- From a scientific perspective, theories should guide and support further research in software engineering

THEORY BUILDING BLOCKS
(Sjøberg et al., 2008)

- Constructs
- Propositions
- Explanations
- Scope

Theories in Software Engineering



Sjøberg, D.I., Dybå, T., Anda, B.C. and Hannay, J.E., 2008. **Building theories in software engineering.** In Guide to advanced empirical software engineering (pp. 312-336). Springer, London.

Theory-Driven Design

> *"Theory building and evaluation can guide the design and analysis of surveys, and surveys can also be applied to test theories."*
>
> *(Wagner et al., 2020)*

- **Theory building** and **survey research** are strongly interrelated
- Initial theories can be drawn from **observations** and available **literature**

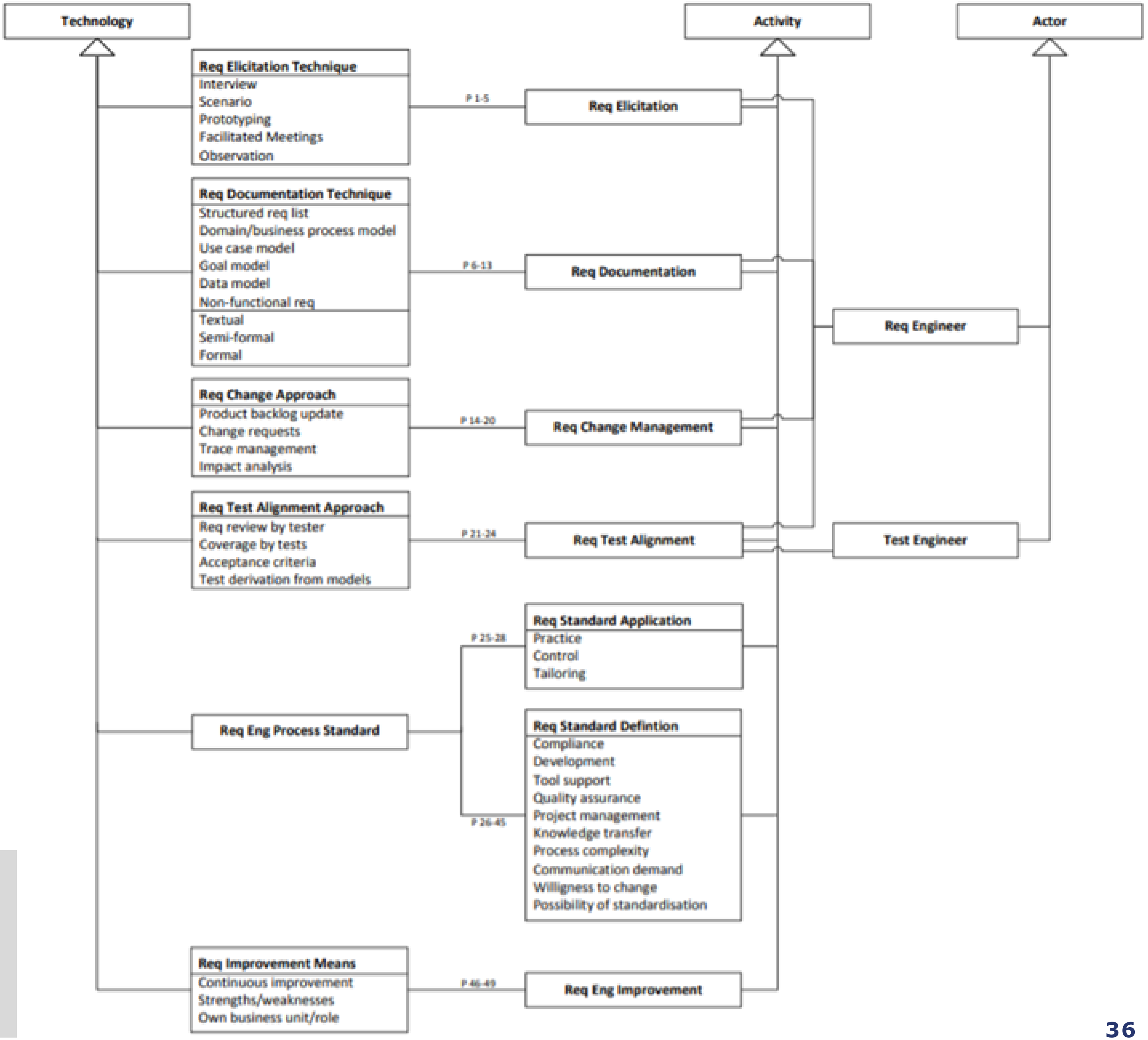
- An initial theory may be a taxonomy of constructs or a set of statements relating constructs
  - For NaPiRE, a set of constructs and propositions was elaborated based on available literature and expert knowledge,
  - For Pandemic Programming, a theoretical model was designed based on related work
  - The surveys, in both cases, were designed to test the theory (and to potentially extend it)


EXAMPLE

Theory-Driven Design Example: NaPiRE

INITIAL THEORY

Constructs		Type
C 1	Requirements Elicitation	Activity
C 2	Requirements Documentation	Activity
C 3	Requirements Change Management	Activity
C 4	Requirements Test Alignment	Activity
C 5	Requirements Standard Application	Activity
C 6	Requirements Standard Definition	Activity
C 7	Requirements Engineering Improvement	Activity
C 8	Requirements Engineer	Actor
C 9	Test Engineer	Actor
C 10	Requirements Elicitation Technique	Technology
C 11	Requirements Documentation Technique	Technology
C 12	Requirements Change Approach	Technology
C 13	Requirements Test Alignment Approach	Technology
C 14	Requirements Engineering Process Standard	Technology
C 15	Requirements Improvement Means	Technology
Scope		
	The theory is supposed to be applicable to contemporary requirements engineering in practice world-wide. There could be differences in different regions of the world because of cultural differences or different economic environments as well as differences in different application domains.	



Wagner, S. *et al*. **Status Quo in Requirements Engineering: A Theory and a Global Family of Surveys.** ACM Transactions on Software Engineering and Methdology, 28(2): 9:1-9:48. 2019.

Theory-Driven Design Example: NaPiRE

No.	Propositions
P 1	Requirements are elicited via interviews
P 2	Requirements are elicited via scenarios
P 3	Requirements are elicited via prototyping
P 4	Requirements are elicited via facilitated meetings (including workshops)
P 5	Requirements are elicited via observation

No.	Explanations	Propositions
E 1	Interviews, scenarios, prototyping, facilitated meetings, and observations allow the requirements engineers to include many different viewpoints including those from nontechnical stakeholders	P1–P5
E 2	Prototypes and scenarios promote a shared understanding of the requirements among stakeholders	P2, P3

Wagner, S. *et al*. **Status Quo in Requirements Engineering: A Theory and a Global Family of Surveys.** ACM Transactions on Software Engineering and Methdology, 28(2): 9:1-9:48. 2019.

Theory-Driven Design Example: NaPiRE

DESIGNED QUESTIONNAIRE

RQ 1 How are requirements elicited and documented?

RQ 2 How are requirements changed and aligned with tests?

RQ 3 How are RE standards applied and tailored?

RQ 4 How is RE improved?

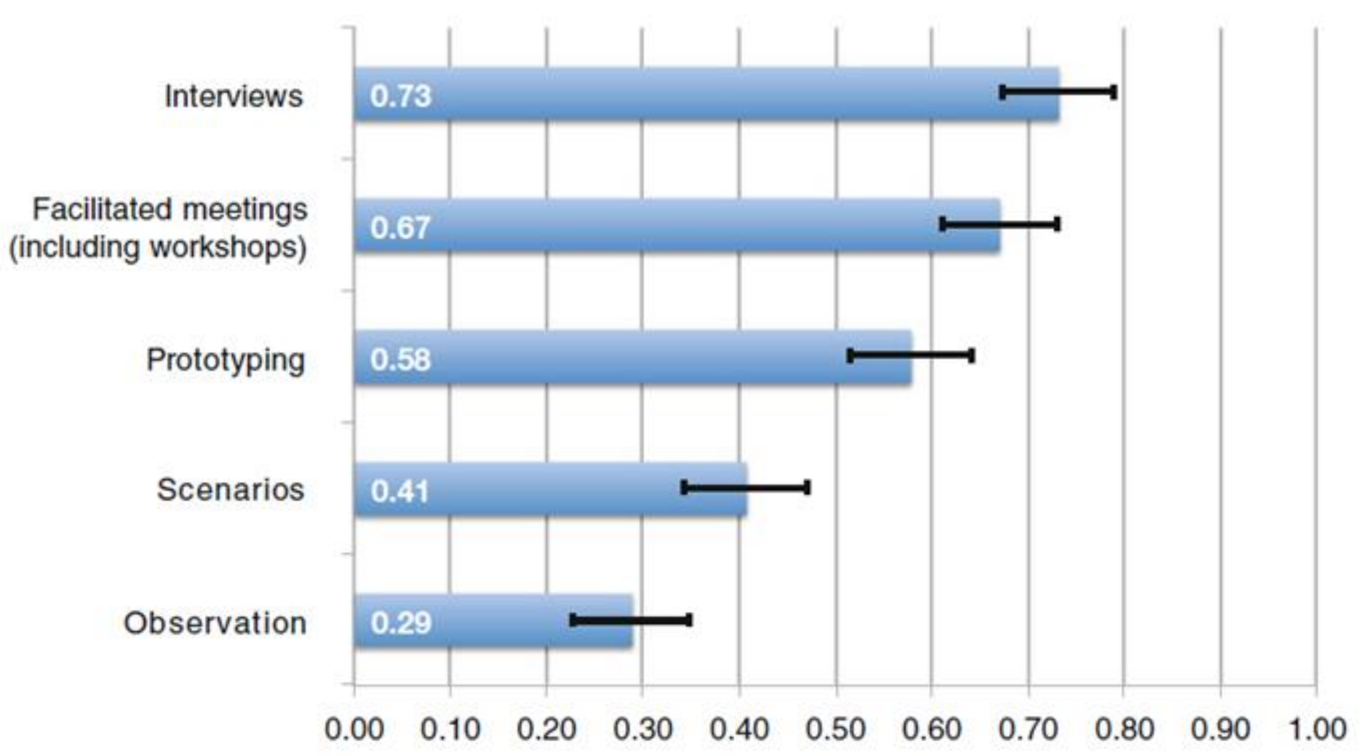
RQ	No.	Question	Type
–	Q 1	What is the size of your company?	Closed(SC)
	Q 2	Please describe the main business area and application domain.	Open
	Q 3	Does your company participate in globally distributed projects?	Closed(SC)
	Q 4	In which country are you personally located?	Open
	Q 5	To which project role are you most frequently assigned?	Closed(SC)
	Q 6	How do you rate your experience in this role?	Closed(SC)
	Q 7	Which organisational role does your company take most frequently in your projects?	Closed(SC)
	Q 8	Which process model do you follow (or a variation of it)?	Closed(MC)
RQ 1	Q 9	How do you elicit requirements?	Closed(MC)
	Q 10	How do you document functional requirements?	Closed(MC)
	Q 11	How do you document non-functional requirements?	Closed(SC)
RQ 2	Q 21	How do you perform change management in your requirements engineering?	Closed(MC)
	Q 12	How do you deal with changing requirements after the initial release?	Closed(SC)
	Q 13	Which traces do you explicitly manage?	Closed(MC)
	Q 14	How do you analyse the effect of changes to requirements?	Closed(MC)
	Q 15	How do you align the software test with the requirements?	Closed(MC)
RQ 3	Q 16	What RE standard have you established at your company?	Closed(MC)
	Q 17	Which reasons do you agree with as a motivation to define a company standard for RE in your company?	Likert
	Q 18	Which reasons do you see as a barrier to define a company standard for RE in your company?	Likert
	Q 19	Is the requirements engineering standard mandatory and practised?	Closed(SC)
	Q 20	How do you check the application of your requirements engineering standard?	Closed(MC)
	Q 22	How is your RE standard applied (tailored) in your regular projects?	Closed(MC)
RQ 4	Q 23	Is your RE continuously improved?	Closed(SC)
	Q 24	Why do you continuously improve your requirements engineering?	Closed(MC)

Wagner, S. *et al*. **Status Quo in Requirements Engineering: A Theory and a Global Family of Surveys.** ACM Transactions on Software Engineering and Methdology, 28(2): 9:1-9:48. 2019.

Theory-Driven Design Example: NaPiRE

No.	Propositions
P 1	Requirements are elicited via interviews
P 2	Requirements are elicited via scenarios
P 3	Requirements are elicited via prototyping
P 4	Requirements are elicited via facilitated meetings (including workshops)
P 5	Requirements are elicited via observation

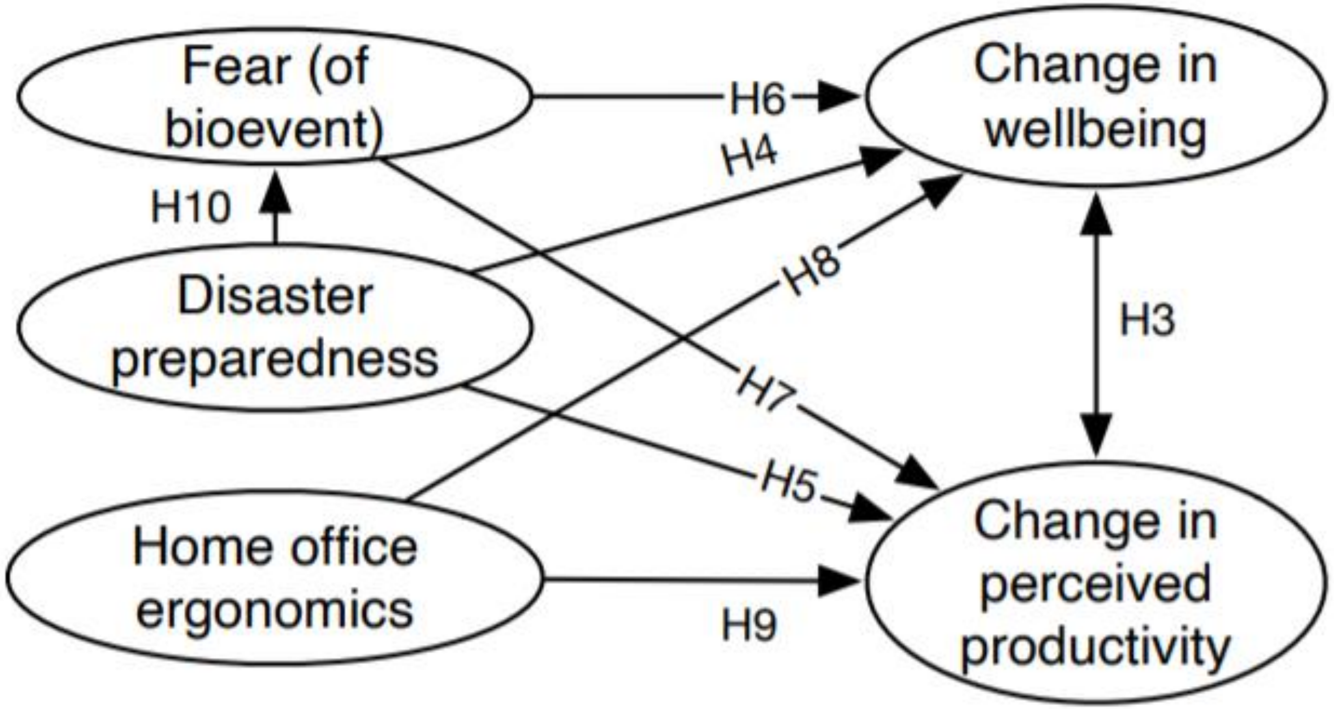
No.	Explanations	Propositions
E 1	Interviews, scenarios, prototyping, facilitated meetings, and observations allow the requirements engineers to include many different viewpoints including those from nontechnical stakeholders	P1–P5
E 2	Prototypes and scenarios promote a shared understanding of the requirements among stakeholders	P2, P3



Wagner, S. *et al*. **Status Quo in Requirements Engineering: A Theory and a Global Family of Surveys.** ACM Transactions on Software Engineering and Methdology, 28(2): 9:1-9:48. 2019.

Theory-Driven Design Example: Pandemic Programming

INITIAL THEORY



Ralph, P., Baltes, S., Adisaputri, G., Torkar, R., Kovalenko, V., Kalinowski, M., Novielli, N., Yoo, S., Devroey, X., Tan, X., Zhou, M., Turhan, B., Hoda, R., Hata, H., Robles, G., Fard, A. M., and Alkadhi, R, **Pandemic Programming How COVID-19 affects software developers and how their organizations can help**. Empirical Software Engineering (2020), 25: 4927-4961. 2020.

SELECTING VALIDATED SCALES FOR THE CONSTRUCTS

Change in wellbeing
We used the WHO's five-item wellbeing index (WHO-5)

Fear (of bioevent)
We adapted the Bracha-Burkle Fear and Resilience (FR) checklist, a triage tool for assessing patients' reactions to bioevents (including pandemics).

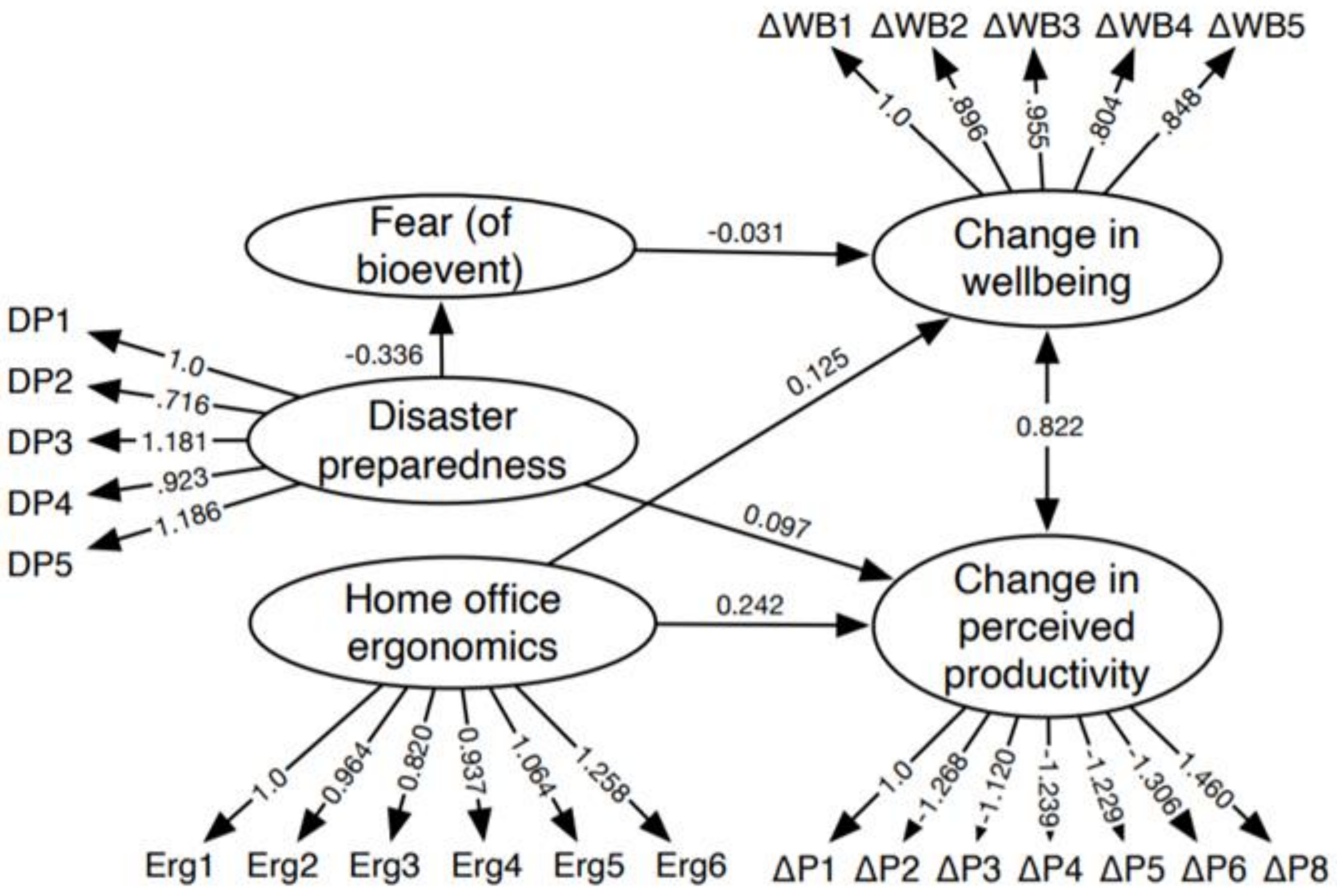
Change in perceived productivity
We used items from the WHO's Health and Work Performance Questionnaire (HPQ)

Home office ergonomics
We could not find a reasonable scale. Based on our reading of the ergonomics literature, we made a simple six-item, six-point Likert scale concerning distractions, noise, lighting, temperature, chair comfort and overall ergonomics.

Disaster preparedness
We adapted Yong et al.'s (2017) individual disaster preparedness scale

Theory-Driven Design Example: Pandemic Programming

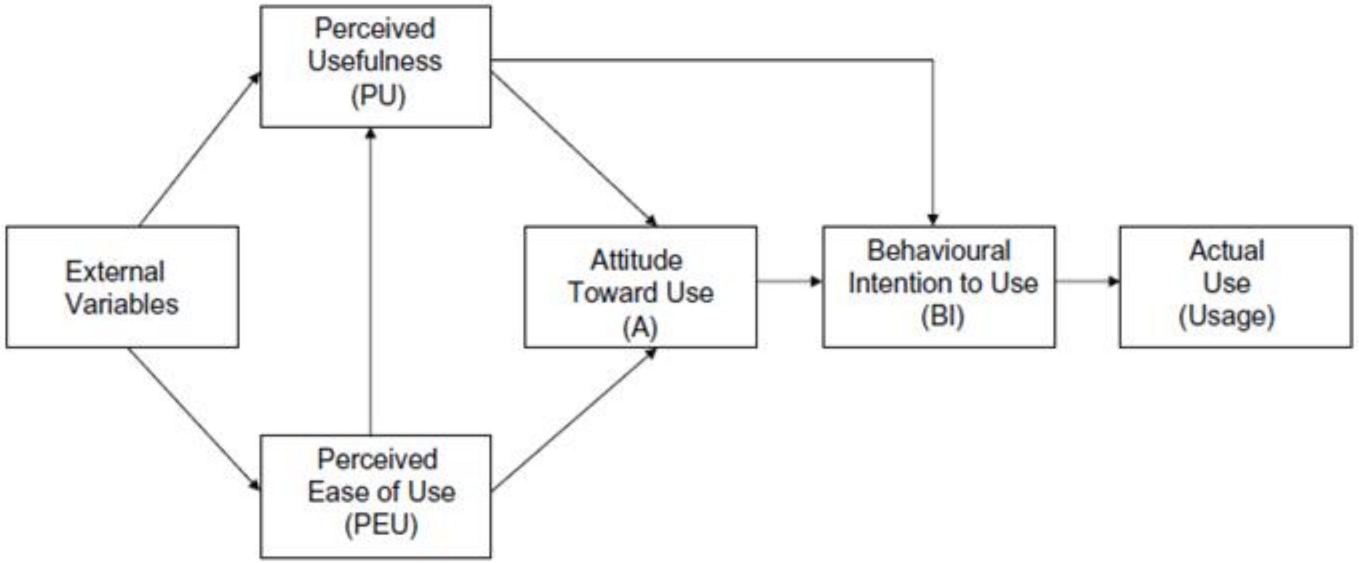
SUPPORTED MODEL



Ralph, P., Baltes, S., Adisaputri, G., Torkar, R., Kovalenko, V., Kalinowski, M., Novielli, N., Yoo, S., Devroey, X., Tan, X., Zhou, M., Turhan, B., Hoda, R., Hata, H., Robles, G., Fard, A. M., and Alkadhi, R, **Pandemic Programming How COVID-19 affects software developers and how their organizations can help**. Empirical Software Engineering (2020), 25: 4927-4961. 2020.

TAM THEORY

The **technology acceptance model** (**TAM**) is an information systems theory that models how users come to accept and use a technology.



The original TAM model (Davis,1989 *apud* Turner *et al.*, 2010).

Criticisms of TAM as a "theory" include its limited explanatory and predictive power, triviality, and lack of practical value

Turner, M., Kitchenham, B., Brereton, P., Charters, S. and Budgen, D., 2010. **Does the technology acceptance model predict actual use? A systematic literature review**. *Information and software technology*, *52*(5), pp.463-479.

TAM QUESTIONNAIRE

The variables within the TAM are typically measured using a short, multiple-item questionnaire

The basic TAM Questionnaire

The specific name of the technology (e.g. the intranet) would replace "*the technology*" in a specific questionnaire.

Responses to statements are given on a Likert-like scale.

Perceived Usefulness Statements

Using *the technology* would improve my performance in doing my job
Using *the technology* at work would improve my productivity
Using *the technology* would enhance my effectiveness in my job
I would find *the technology* useful in my job

Perceived Ease of Use Statements

Learning to operate *the technology* would be easy for me
I would find it easy to get *the technology* to do what I want it to do
It would be easy for me to become skilful in the use of *the technology*
I would find *the technology* easy to use

Behavioural Intention to use

I intend to *use the technology* regularly at work

Turner, M., Kitchenham, B., Brereton, P., Charters, S. and Budgen, D., 2010. **Does the technology acceptance model predict actual use? A systematic literature review**. *Information and software technology*, *52*(5), pp.463-479.

Alternatives for technology acceptance …

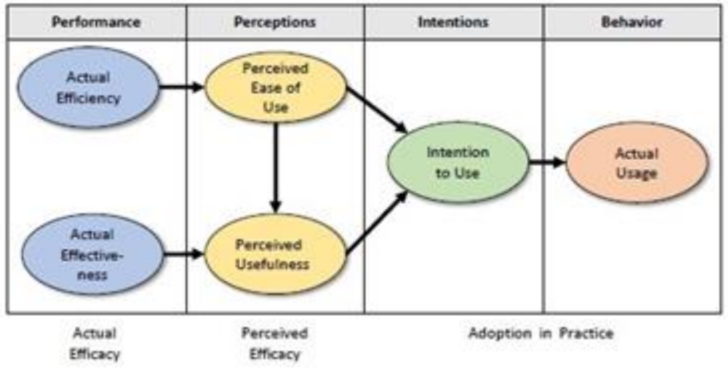
Unified theory of acceptance and use of technology (Venkatesh et al., 2003) *

Technology Acceptance Model (Davis, 1989) **

Matching Person & Technology model (Sherer, 1986)

Diffusion of Innovations Theory (Rogers, 1962)

Method Evaluation Model (Moody, 2003) …



* Venkatesh, V., Morris, M.G., Davis, G.B. and Davis, F.D., 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly*, pp.425-478. 35.000+ citations by 2021

** Davis, F.D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pp.319-340. 63.000+ citatins by 2021

Evaluating Theories

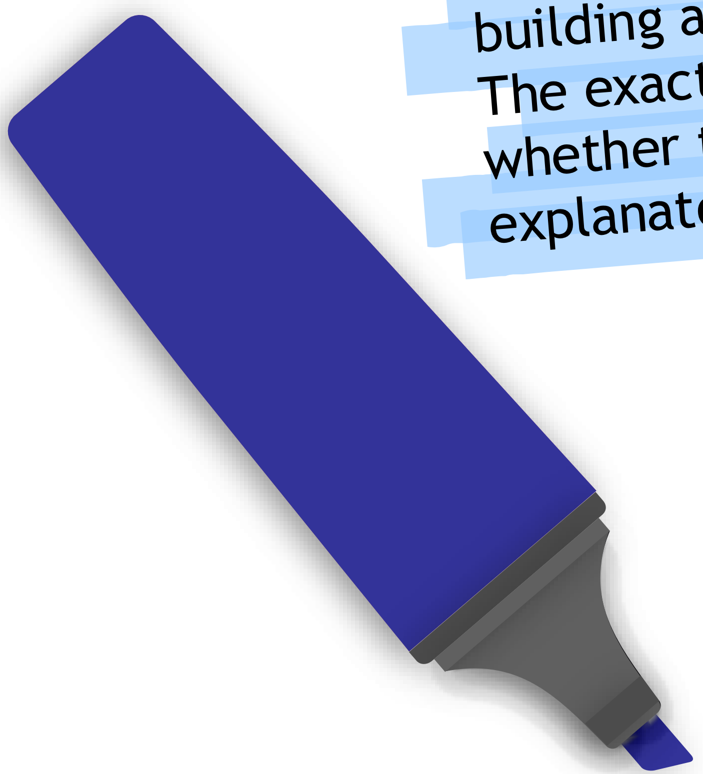
Table 1 Criteria for evaluating theories

Testability	The degree to which a theory is constructed such that empirical refutation is possible
Empirical support	The degree to which a theory is supported by empirical studies that confirm its validity
Explanatory power	The degree to which a theory accounts for and predicts all known observations within its scope, is simple in that it has few ad hoc assumption, and relates to that which is already well understood
Parsimony	The degree to which a theory is economically constructed with a minimum of concepts and propositions
Generality	The breadth of the scope of a theory and the degree to which the theory is independent of specific settings
Utility	The degree to which a theory supports the relevant areas of the software industry

Sjøberg, D.I., Dybå, T., Anda, B.C. and Hannay, J.E., 2008. **Building theories in software engineering.** In Guide to advanced empirical software engineering (pp. 312-336). Springer, London.

Survey Research and Theory Building

Key Takeaways *(Wagner et al., 2020)*:



A

Survey research and theory building are strongly interrelated. The exact relationship depends on whether the theory is descriptive, explanatory, or predictive.

B

Theories are of high value to guide the design of surveys.

C

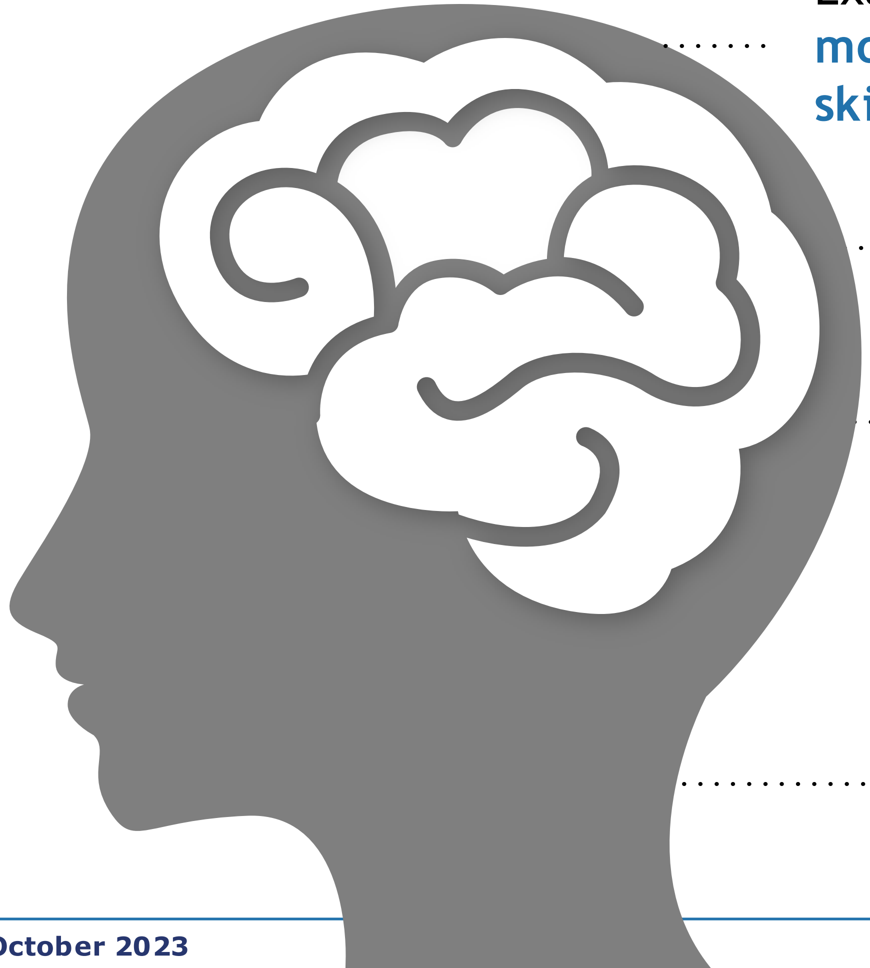
Survey data supports the definition or refinement of constructs, relationships, explanations, and the scope of a theory as well as testing of a theory.

D

Use validated scales as much as possible to improve construct validity.

Issues When Assessing Psychological Constructs

Psychological constructs are theoretical concepts to model and understand human behavior, cognition, affect, and knowledge *(Binning, 2016)*



Examples include **happiness**, **job satisfaction**, **motivation**, **commitment**, **personality**, **intelligence**, **skills**, and **performance**

These constructs can only be assessed **indirectly**

We need ways to **proxy** our **measurement** of a construct in robust, valid, and reliable ways

- This is why, whenever we wish to investigate psychological constructs and their variables, we need to either develop or adopt measurement instruments that are **psychometrically validated**

Scientists have investigated issues of **validity** and **reliability** of psychological tests

Validity and Reliability in Psychometrics *(AERA et al., 2014)*

VALIDITY

- ✓ The degree to which evidence and theory support the interpretation of test scores for proposed uses of tests
- ✓ We need to ensure that any meaning we provide to the values obtained by a measurement instrument needs to be validated

RELIABILITY

- ✓ Consistency of a questionnaire score in repeated instances of it; or
- ✓ Coefficient between scores on two equivalent forms of the same test

Characteristics and Purpose

> *Software engineering research should favor psychometric validation of tests.*

(Wagner et al., 2020)

Wagner, S., Mendez, D., Felderer, M., Graziotin, D. and Kalinowski, M., 2020. **Challenges in survey research.** In: Contemporary Empirical Methods in Software Engineering (pp. 93-125). Springer, Cham.

Key Takeaways *(Wagner et al., 2020)*:

A

Representing and assessing constructs on human behavior, cognition, affect, and knowledge is a difficult problem that requires psychometrically validated measurement instruments.

B

Software engineering research should either adopt or develop psychometrically validated questionnaires.

C

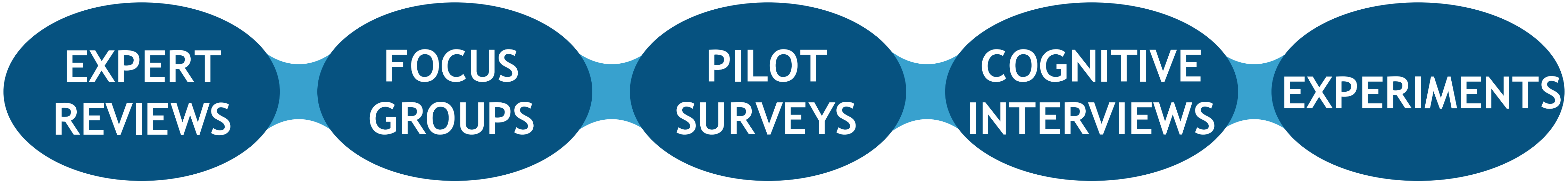
Adoption or development of psychometrically validated questionnaires should consider psychometric validity and reliability issues, which are diverse and very different from the usual and common validity issues we see in "Threats to Validity" sections.

D

Software engineering research should introduce studies on the development and validation of questionnaires.

- Used to assess the **validity** and **reliability** of the survey instrument

- A survey can be evaluated, to avoid **threats to validity and reliability**, using the following methods *(Robson, 2002 apud Linaker et al., 2015)*:



SAMPLING

Sampling

- At the beginning of any design of survey research, we should clarify what the **target population** is that we try to characterize and generalize to
  - Statistical analysis relies on systematic sampling from this target population

- In software engineering surveys, the unit of analysis that defines the granularity of the target population is often *(de Mello et al. 2015)*

AN ORGANIZATION

A SOFTWARE TEAM OR PROJECT

AN INDIVIDUAL

Sampling

- For common research questions, we are typically interested in producing results related to all organizations that develop software **in the world** or all software developers **in the world**
  - We want to find theories that have a scope as wide as possible

- We have no solid understanding about the **target population**
  - Which companies are developing software?
  - How many software developers are there in the world?
  - What are the demographics of software engineers in the world?

- We face enormous difficulties to discuss **representativeness** of a sample, the **needed size** of the sample and, therefore, to what degree we can **generalize** our results

Representativeness

- Scientists often rely on **demographic information** published by governmental or other public bodies such as statistical offices
  - These bodies are, so far, rather unhelpful for our task, because they do not provide a good idea about software-developing companies

- There are possibilities to approach the demographics of software engineers
  - Commercial **providers of data** from large surveys such as Evans Data Corporation
    - Estimated number of developers worldwide as of 2018: 23 million
    - Include information on different roles, genders, used development processes and technologies
  - An open alternative is the ***Stack Overflow Annual Developer Survey***

Representativeness

- Having demographic information, we can design our survey in a way that we collect **comparable data**

- Then, we can **compare the distributions** in our survey and the larger surveys to estimate **representativeness**

A Should be part of the interpretation and discussion of the results

B Prevents us from overclaiming

C Gives us more credibility in case we cover the population well

Sample Size Estimation

- A good sample size (n) can be estimated as follows *(Yamane, 1973 apud Wagner et al., 2020)*:

$$n = \frac{N}{1 + Ne^2}$$

n - sample size

N - population size

e - level of precision (often set to 0.05 or 0.01)

- Reasonable sample size for software developers (using precision 0.05):

$$n = \frac{23{,}000{,}000}{1 + 23{,}000{,}000 \cdot 0.05^2} = 400$$

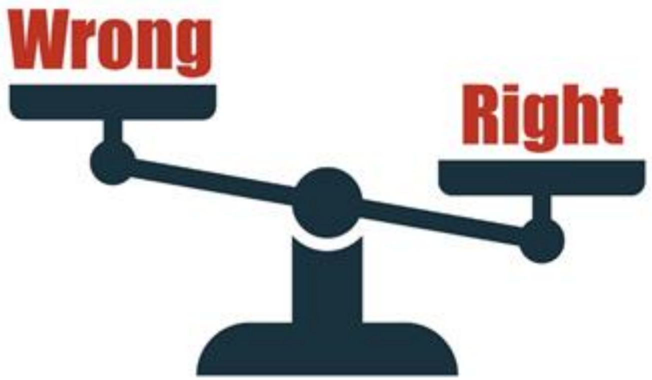
Ethics

- In software engineering, there is yet **no established standard or guidelines** on how to conduct surveys ethically

- The *Insight Association* provides ethical guidelines that consider **unethical sampling**, among other practices: *"Collection of respondent emails from Websites, portals, Usenet or other bulletin board postings without specifically notifying individuals that they are being 'recruited' for research purposes."*

- We will probably need **flexible rules and guidelines** to keep developers in social media from being spammed by study requests while still allowing research to take place

- We should all consider thoughtfully **how and whom** we contact for a survey study



Key Takeaways:

A

There is no suitable official data on the number and properties of software developing companies in the world

B

Ethics needs to be considered before contacting potential survey participants

C

For individual software engineers, existing demographic studies can be used to assess a survey's representativeness

D

For the estimate of 23 million developers worldwide, a good sample size would be 400 respondents

DATA COLLECTION

Data Collection - Invitation and Follow-up

Strategies to approach the population:

CLOSED INVITATIONS

- ✓ approaching known groups or individuals to participate per invitation-only
- ✓ restricting the survey access to those invited

OPEN INVITATIONS

- ✓ approaching a broader, often anonymous audience via open survey access
- ✓ anyone with a link to the survey can participate

Invitation and Follow-up

The strategy has implications on the survey design and the recruitment approaches

CLOSED INVITATIONS

- ✓ allows to accurately choose the respondents based on predefined characteristics and the suitability to provide the required information
- ✓ allows to accurately calculate the response rate and control the participation along the data collection
- ✓ typically implies in a lower number of total responses

OPEN INVITATIONS

- ✓ allows to spread the invitation broader, e.g., via public forums, mailing lists, social media, or at venues of conferences and workshops.
- ✓ does not require to carefully select lists of subject candidates and to approach them individually, but it also comes at the cost of loosing control in who provides the responses
- ✓ requires defining to define proper demographic questions that allow analyzing the extent to which the respondents are suitable to provide the required information

Key Takeaways *(Wagner et al., 2020)*:

A

Both strategies to approach the target population (closed and open invitations) can be applied, but have distinct implications on the survey design and the recruitment approaches.

B

Closed invitations are suitable in situations in which it is possible to precisely identify and approach a well-defined sample of the target population. They may also be required in situations where filtering out participants that are not part of the target population would be difficult, harming the sample representativeness.

C

Open invitations allow reaching out for larger samples. However, they typically require more carefully considering context factors when designing the survey instruments. These context factors can then be used during the analyses to filter out participants that are not representative (e.g., applying the blocking principle to specific context factors).

STATISTICAL ANALYSIS

- DESCRIPTIVE STATISTICS
- NULL-HYPOTHESIS SIGNIFICANCE TESTING
- BOOTSTRAPPING CONFIDENCE INTERVALS
- BAYESIAN ANALYSIS
- STRUCTURAL EQUATION MODELING

With the often large number of participants in surveys, we usually aim at a statistical analysis of the survey results

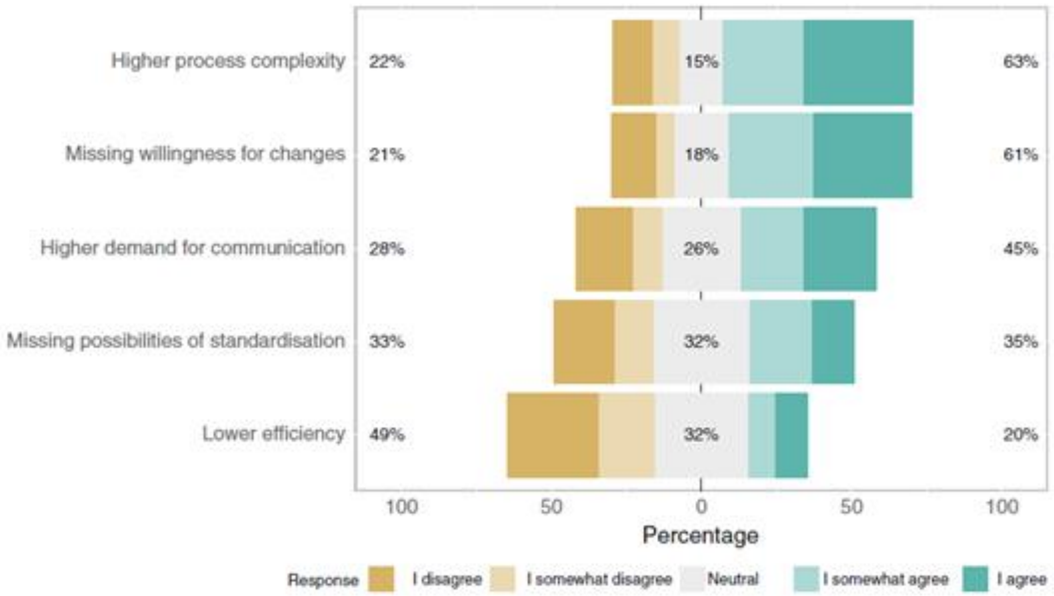
A majority of the questionnaires are typically composed of closed questions that have quantitative results

Descriptive Statistics

- The goal of descriptive statistics is to **characterize** the answers to one or more questions of our specific sample
- We do not yet talk about generalizing to the population
- Which descriptive statistic is suitable depends on **what we are interested** in most and the **scale** of the data

Scale	Nominal	Ordinal	Interval	Ratio
Values Counting	X	X	X	X
Values Ordering		X	X	X
Equidistant Intervals			X	X
Values Division				X

- Descriptive statistics for ordinal scales (e.g., Likert scales)
  - Frequency counting, mode, median, minimum, maximum, median absolute deviation (MAD), interquartile range (IQR)
  - An interesting alternative is showing the whole distribution of ordinal data in a stacked bar chart



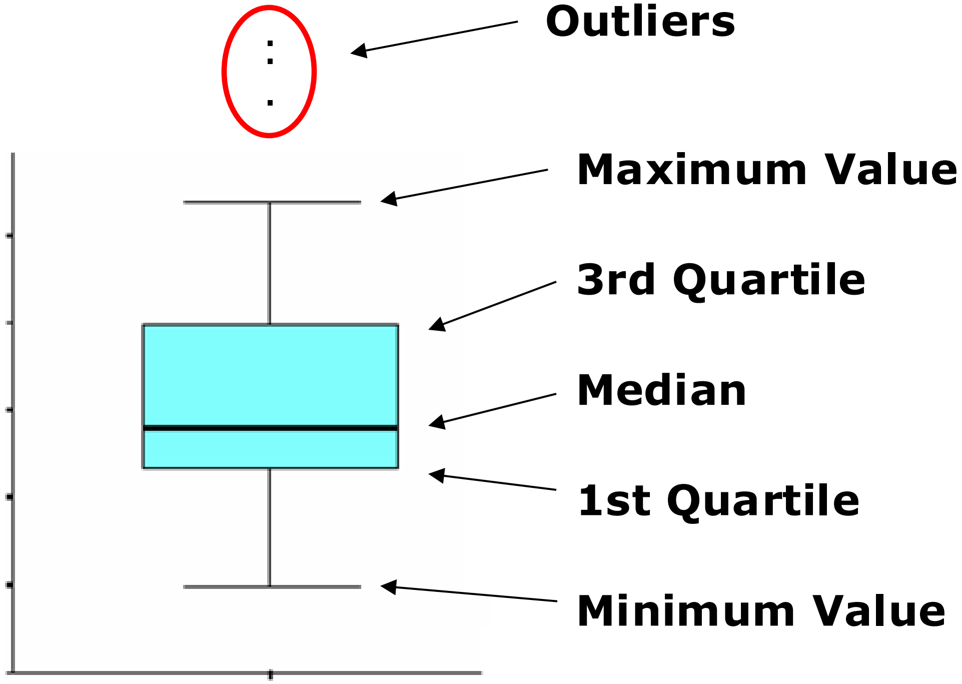
Wagner et al., 2019

Generated using the Likert package in R

http://www.labape.com.br/rprimi/statR/T7_plus_likert.html

Descriptive Statistics

- For **interval** or **ratio** scales we can use all available descriptive statistics, such as mean, variance, and standard deviation.
- Still, we recommend using **boxplots**, to enable eliminating outliers by using the quartile method



Quartile Method

Lower Outliers: Q1 - 1.5*IQR
Upper Outliers: Q3 + 1.5*IQR
Where IQR = Q3 – Q1.

Null-hypothesis Significance Testing (NHST)

Descriptive statistics concern the **sample**
Inference statistics concern the **population**

We need hypotheses to evaluate

- ✓ A survey should be guided by a theory
- ✓ Propositions can be operationalized into hypotheses to test with the survey data

In surveys we typically have:

- ✓ Point estimate hypotheses for answers to single questions
- ✓ Hypotheses on correlations between answers to two questions

Null-hypothesis Significance Testing (NHST)

- In general, two hypotheses are defined:
  - **Null Hypothesis(H0):** indicates the observed differences are coincidental. It means that this is the hypothesis the researcher would like most to reject with high confidence
  - **Alternative Hypothesis(H1):** represents the hypothesis indicating some type of effect, that can be accepted, or tested

- Types of Errors
  - **Type I (α):** it happens when the statistical test indicates the existence of a relationship between cause and effect that actually does not exist
  - **Type II (ß):** it happens when the statistical test does not indicate a relationship between cause and effect that actually does exist

- Statistics tests allow confirming or refuting hypotheses (according to a previously defined **significance level** - *a-value*)

Null-hypothesis Significance Testing (NHST)

- Significance Testing
  - Shows the likelihood of an type-I error to happen
    - Most common significance level (α): 10%, 5%, 1% and 0.1%
    - We call ***p-value*** the lowest level of significance that can be used to reject the null hypothesis
    - We say there is statistical significance when the calculated *p-value* is lower than the adopted significance level (*a-value*)

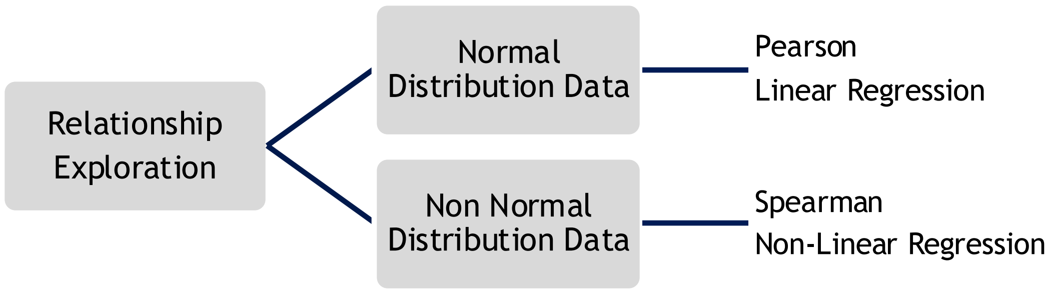
- Besides significance testing, it is important to also look at effect sizes
  - Cohen's ***d*** is defined as the difference between two means divided by a standard deviation for the data

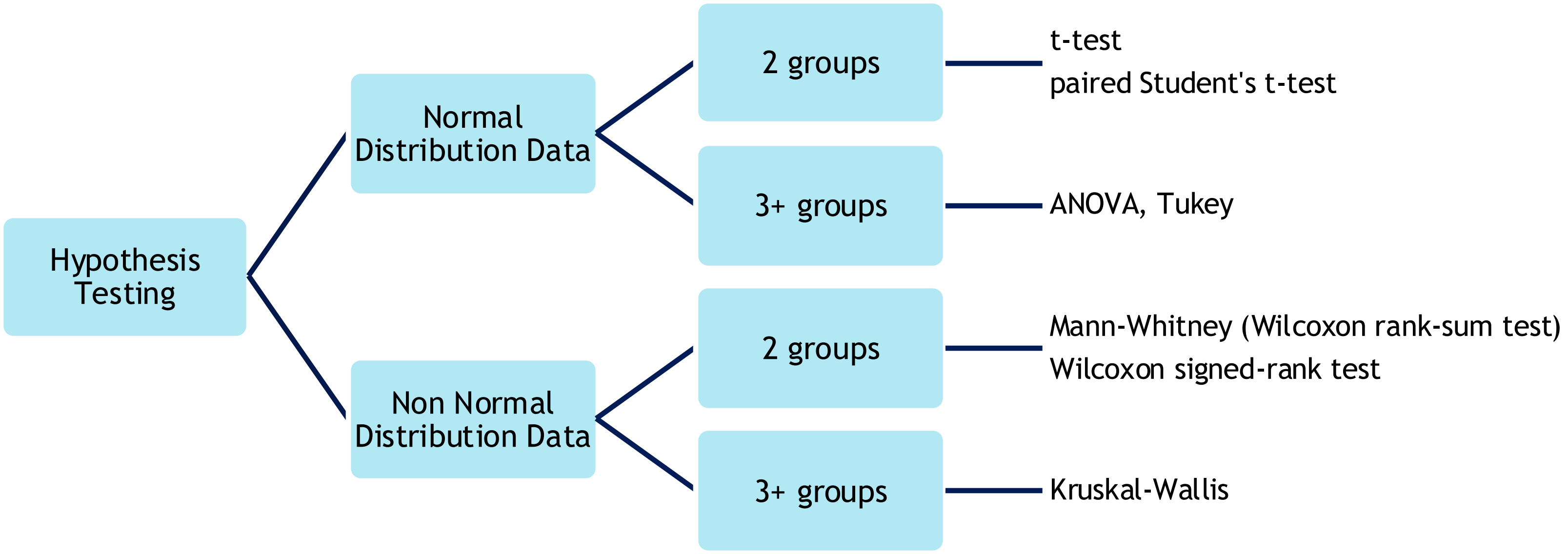
$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Several statistical significance tests can be applied, with differences in their statistical power

(Power= P (H_0 rejected | H_o is false))

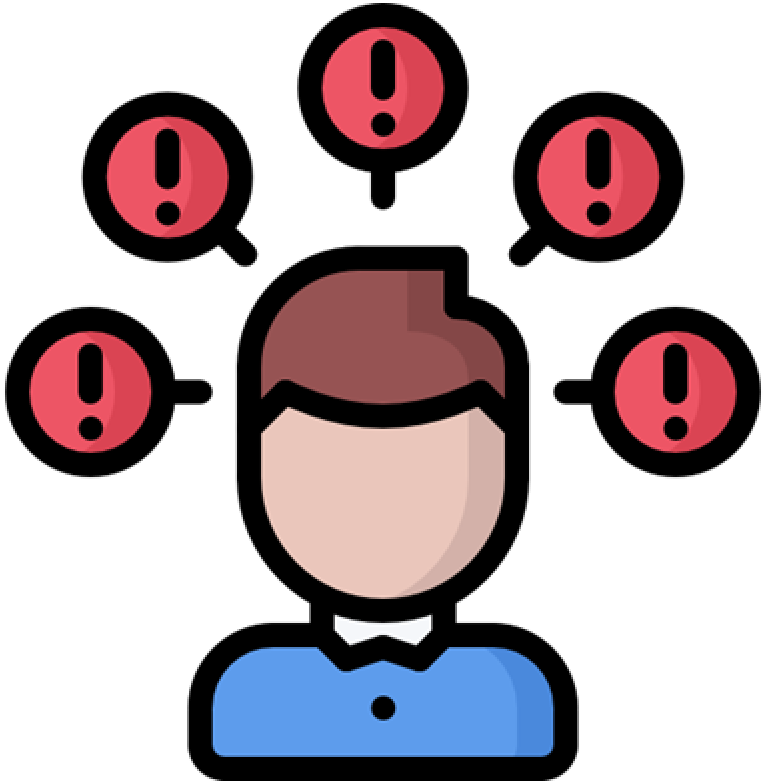
- The statistical test with the highest power shall be used to evaluate the hypotheses





Null-hypothesis Significance Testing (NHST)

- Problems with NHST
  - Dichotomous nature of its results
  - Requires a representative sample of the population, otherwise it is unclear what NHST actually means



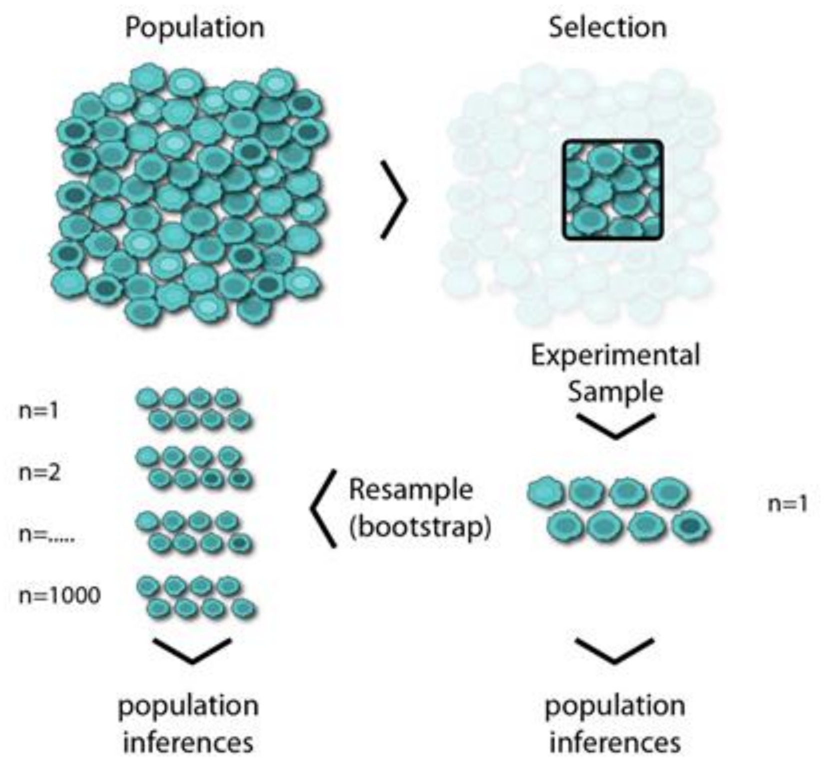
- We need alternatives ...

Bootstrapping Confidence Intervals

- Replaces fixed significance level tresholds

- Involves **estimating a confidence interval** around a metric we are interested in
  - How large is the confidence interval?
  - How strongly do confidence intervals of methods to compare overlap?

- Idea of bootstrapping:
  - We repeatedly take **samples with replacement** and calculate the statistic we are interested in
  - This is repeated a large number of times and, thereby, provides us with an understanding of the distribution of the sample

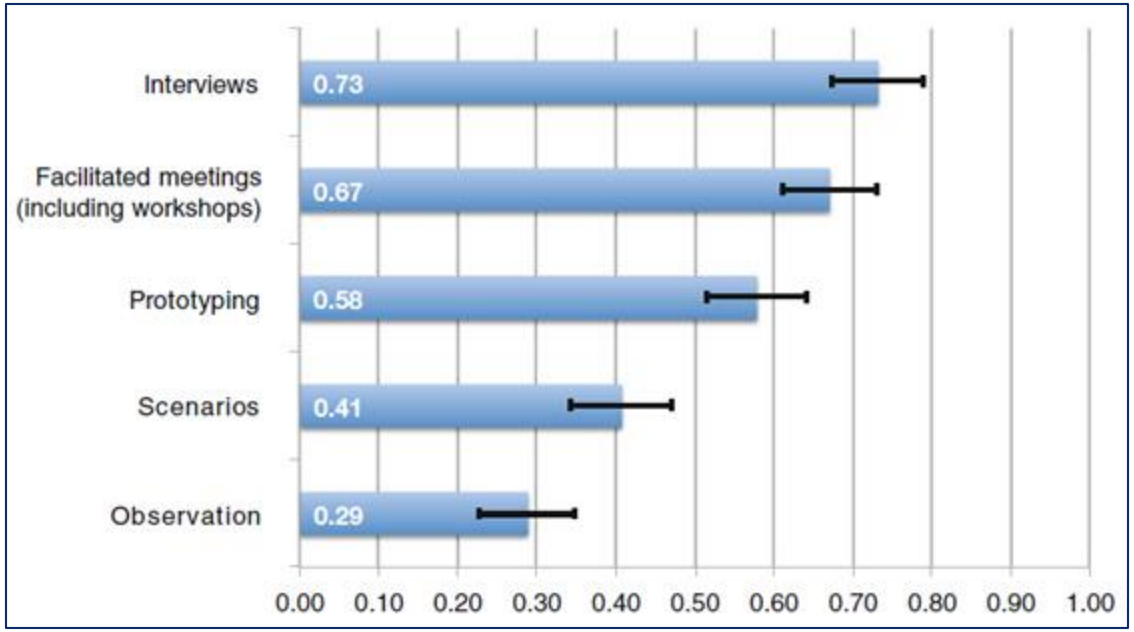


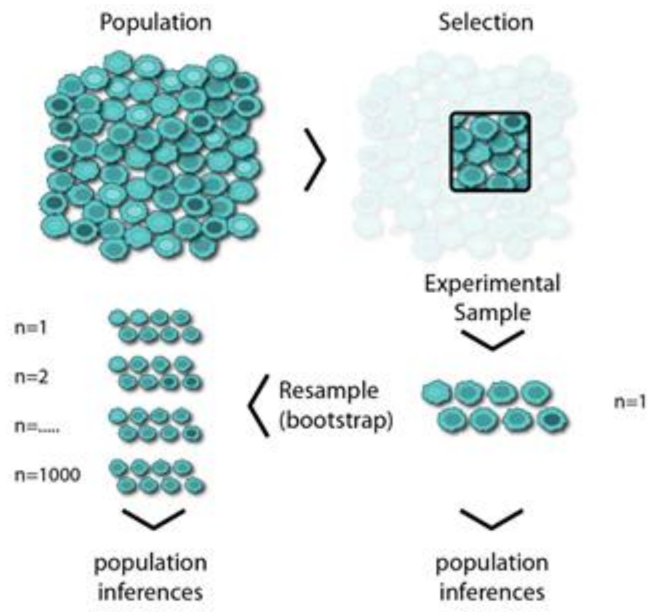
Source: https://medium.com/swlh/bootstrap-sampling-using-pythons-numpy-85822d868977


EXAMPLE

Bootstrapping Confidence Intervals: Example

- 1000 times resampling for bootstrapping confidence intervals

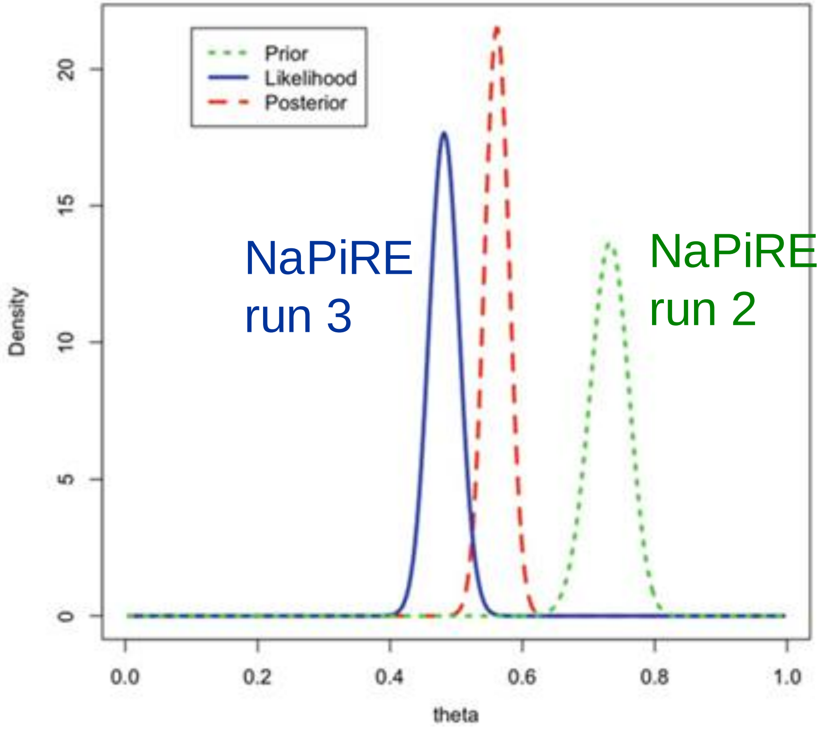




The Bootstrap Assumption: The original sample approximates the population from which it was drawn. So resamples from this sample approximate what we would get if we took many samples from the population. The bootstrap distribution of a statistic, based on many resamples, approximates the sampling distribution of the statistic, based on many samples.

Bayesian Analysis

- In Bayesian statistics, probability is understood as a representation of the state of knowledge or belief
  - Acknowledges uncertainty
  - Allows integrating existing evidence and accumulating knowledge



Workshops for eliciting requirements (Wagner et al., 2020)

Further Reading:

Torkar, R., Feldt, R. and Furia, C.A., 2020. **Bayesian Data Analysis in Empirical Software Engineering: The Case of Missing Data.** In Contemporary Empirical Methods in Software Engineering (pp. 289-324). Springer, Cham.

Structural Equation Modeling

- **Structural Equation Modeling (SEM)** is used to test theories involving constructs (also called latent variables).
  - In our Pandemic Programming survey example fear, disaster preparedness, home office ergonomics, wellbeing and productivity are all constructs

- To design a structural equation model, we first define a **measurement model**, which maps each reflective indicator into its corresponding construct.
  - For example, each of the five items comprising the WHO5 wellbeing scale is modeled as a reflective indicator of wellbeing

- SEM uses **Confirmatory Factor Analysis (CFA)** to estimate each construct as the shared variance of its respective indicators

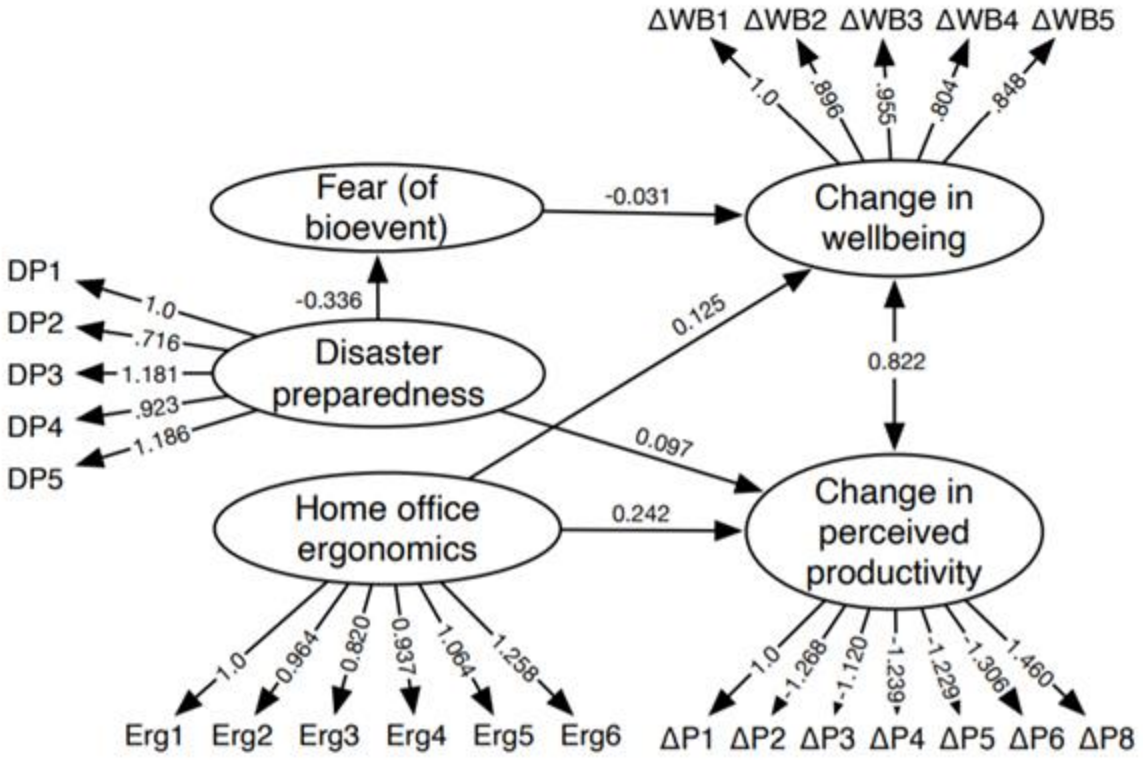
Structural Equation Modeling

- Next, we define the **structural model**, which identifies the expected relationships among the constructs
  - The constructs we are attempting to predict are referred to as **endogenous** (dependent variables), while the predictors are **exogenous** (independent variables)

- SEM uses a **path modeling technique** (e.g. regression) to build a model that predicts the endogenous (latent) variables based on the exogenous variables, and to estimate both the strength of each relationship and the overall accuracy of the model


EXAMPLE

SUPPORTED MODEL

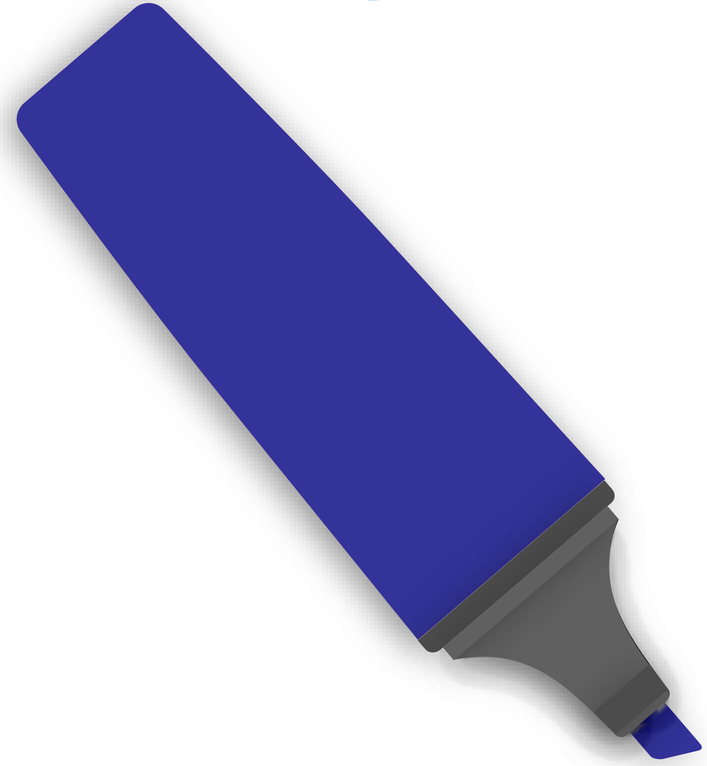


The arrows between the constructs show the supported causal relationships.

The path coefficients (the numbers on the arrows) indicate the relative strength and direction of the relationships.

Ralph, P., Baltes, S., Adisaputri, G., Torkar, R., Kovalenko, V., Kalinowski, M., Novielli, N., Yoo, S., Devroey, X., Tan, X., Zhou, M., Turhan, B., Hoda, R., Hata, H., Robles, G., Fard, A. M., and Alkadhi, R, **Pandemic Programming How COVID-19 affects software developers and how their organizations can help**. Empirical Software Engineering (2020), 25: 4927-4961. 2020.

Key Takeaways *(Wagner et al., 2020)*:



A

Always make clear whether you aim at analyzing opinions or facts

B

Descriptive statistics are always helpful

C

NHST inferential statistics are useful to test theoretical propositions

D

Bootstrapping confidence intervals helps to deal with uncertain sampling

E

Bayesian analysis allows us to directly integrate prior knowledge

F

Structural equation modeling (SEM) is a powerful multivariate analysis technique that is widely used in the social sciences and that should be further used in computer science research

QUALITATIVE ANALYSIS

Qualitative Analysis

Besides the common focus on statistical analysis, surveys can also be **qualitative** and contain **open questions**

Open questions **do not impose restrictions** on respondents and allow them to more precisely describe the phenomena of interest according to their perspective and perceptions

However, they can lead to a **large amount of qualitative data to analyze**, which is not easy and may require a significant amount of resources

The answers to such open questions can help researchers to further understand a phenomenon eventually including causal relations among theory constructs and theoretical explanations

Open questions can help generating new theories

A research method commonly employed to support qualitative analyses is Grounded Theory.

There are at least three main streams of GT:

- ✓ Glaser's GT (classic or Glaserian GT) *(Glaser, 1992)*
- ✓ Corbin and Strauss' GT (Straussian GT) (*Corbin and Strauss, 1990)*
- ✓ Charmaz's constructivist GT *(Charmaz, 2014)*

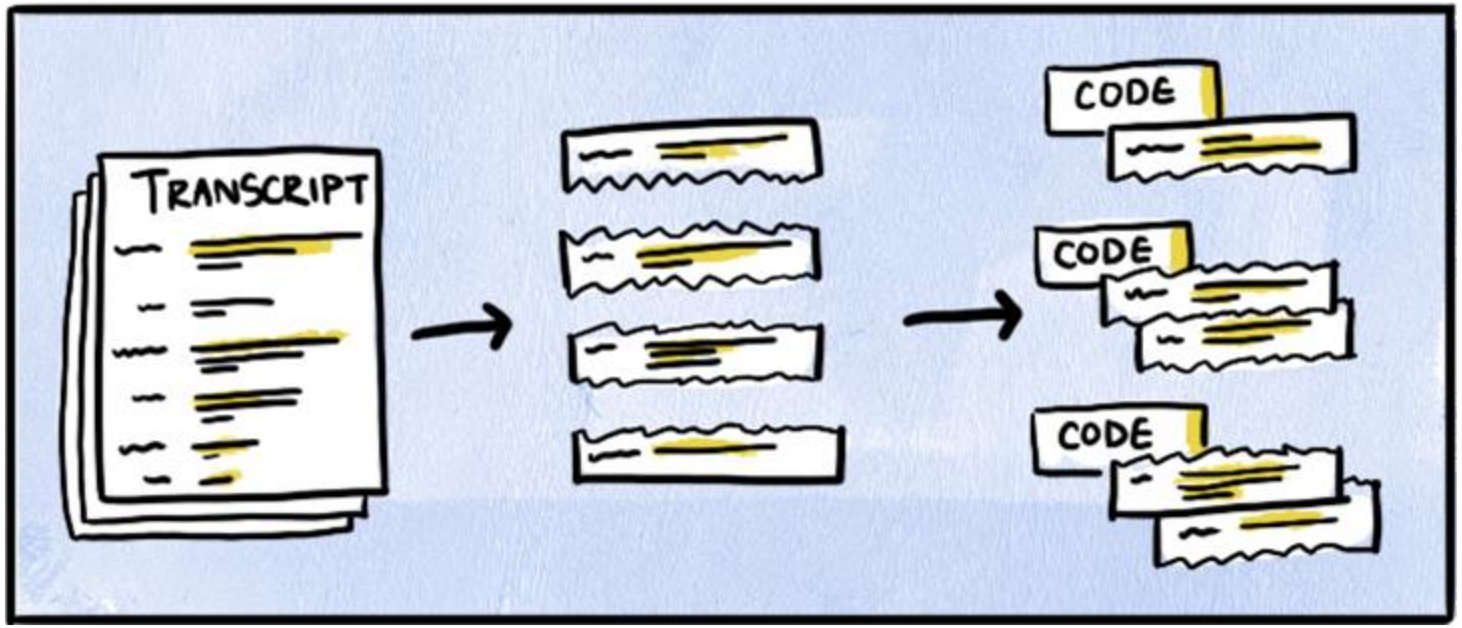
Grounded theory, "in theory", involves inductively generating theory from data.

Few "GT" Studies Generate Theory *(Stol et al., 2016)*

Open Coding

1 Turn your data into small, discrete components of data

2 Code each discrete pieces of data with a descriptive label

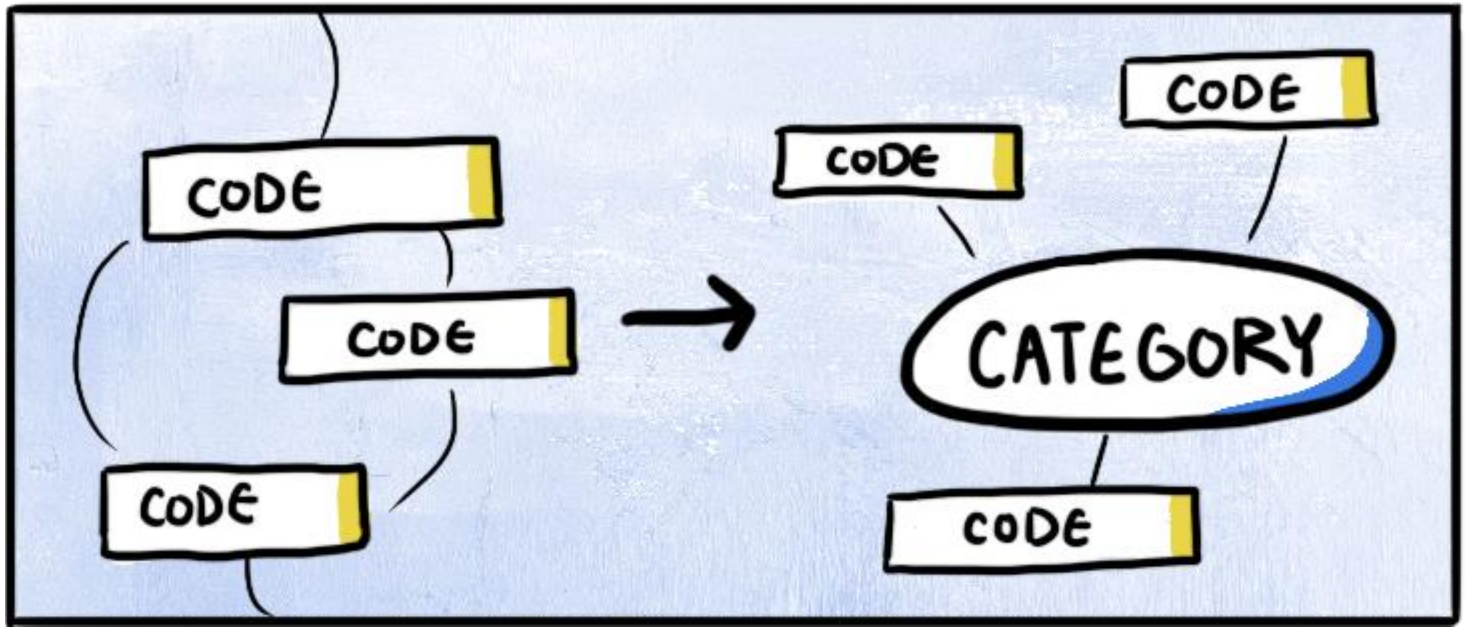


Source: *https://delvetool.com/blog/openaxialselective*

Corbin, J.M. and Strauss, A., 1990. **Grounded theory research: Procedures, canons, and evaluative criteria.** Qualitative sociology, 13(1), pp.3-21.

3 Find connections and relationships between codes

4 Aggregate and condense codes into broader categories



Source: https://delvetool.com/blog/openaxialselective

Corbin, J.M. and Strauss, A., 1990. **Grounded theory research: Procedures, canons, and evaluative criteria.** Qualitative sociology, 13(1), pp.3-21.

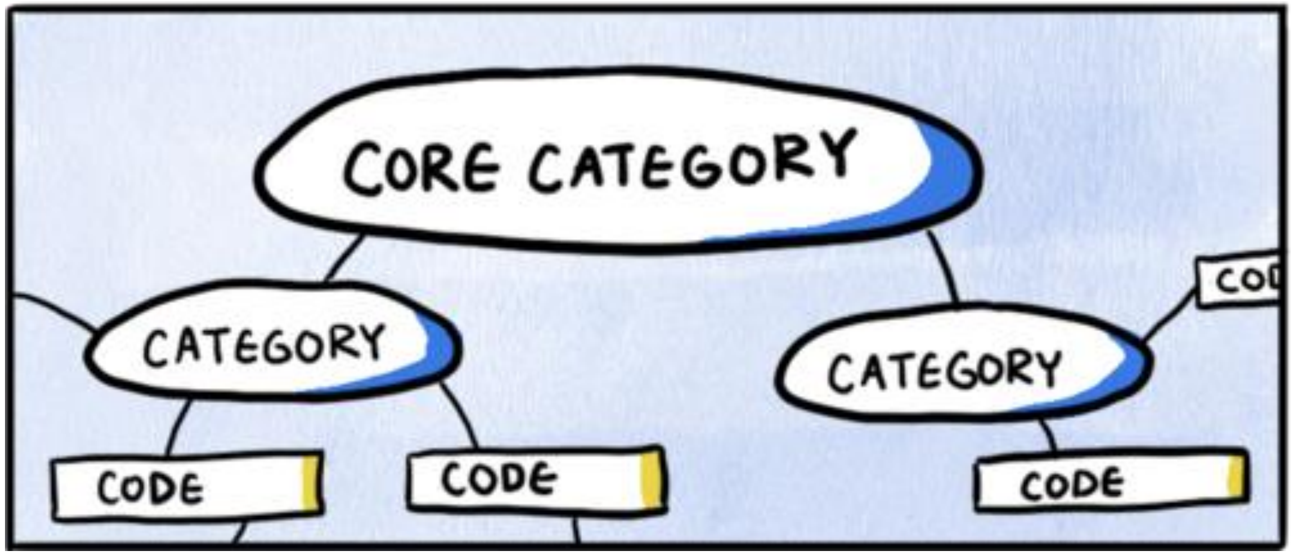
Selective Coding

5 Bring it together with one overarching category

6 Identify the connections between this overarching category and the rest of your codes and data

7 Remove categories or codes that don’t have enough supporting data

8 Read the transcript again, and code according to this overarching category

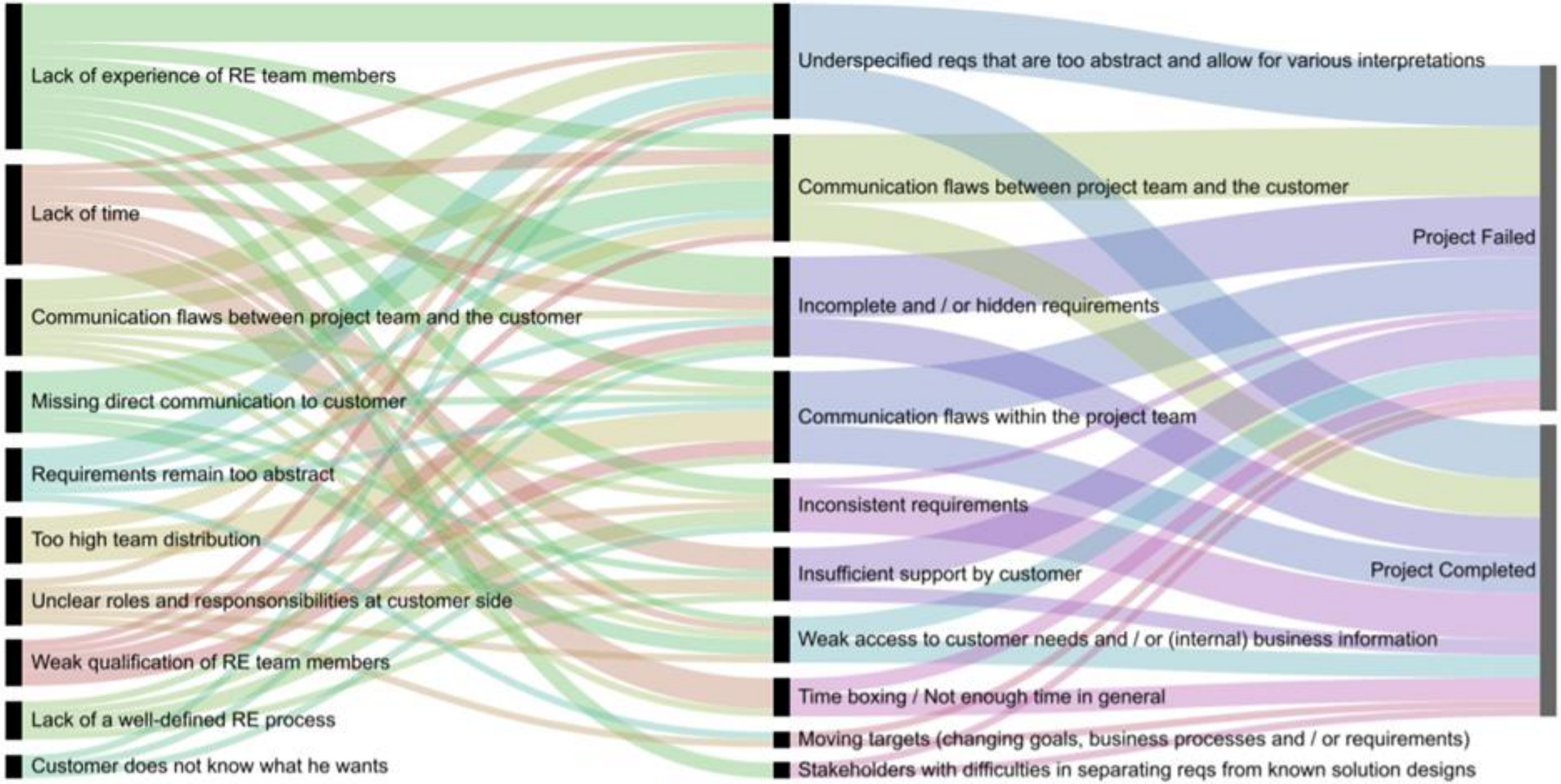


Source: https://delvetool.com/blog/openaxialselective

Corbin, J.M. and Strauss, A., 1990. **Grounded theory research: Procedures, canons, and evaluative criteria.** Qualitative sociology, 13(1), pp.3-21.

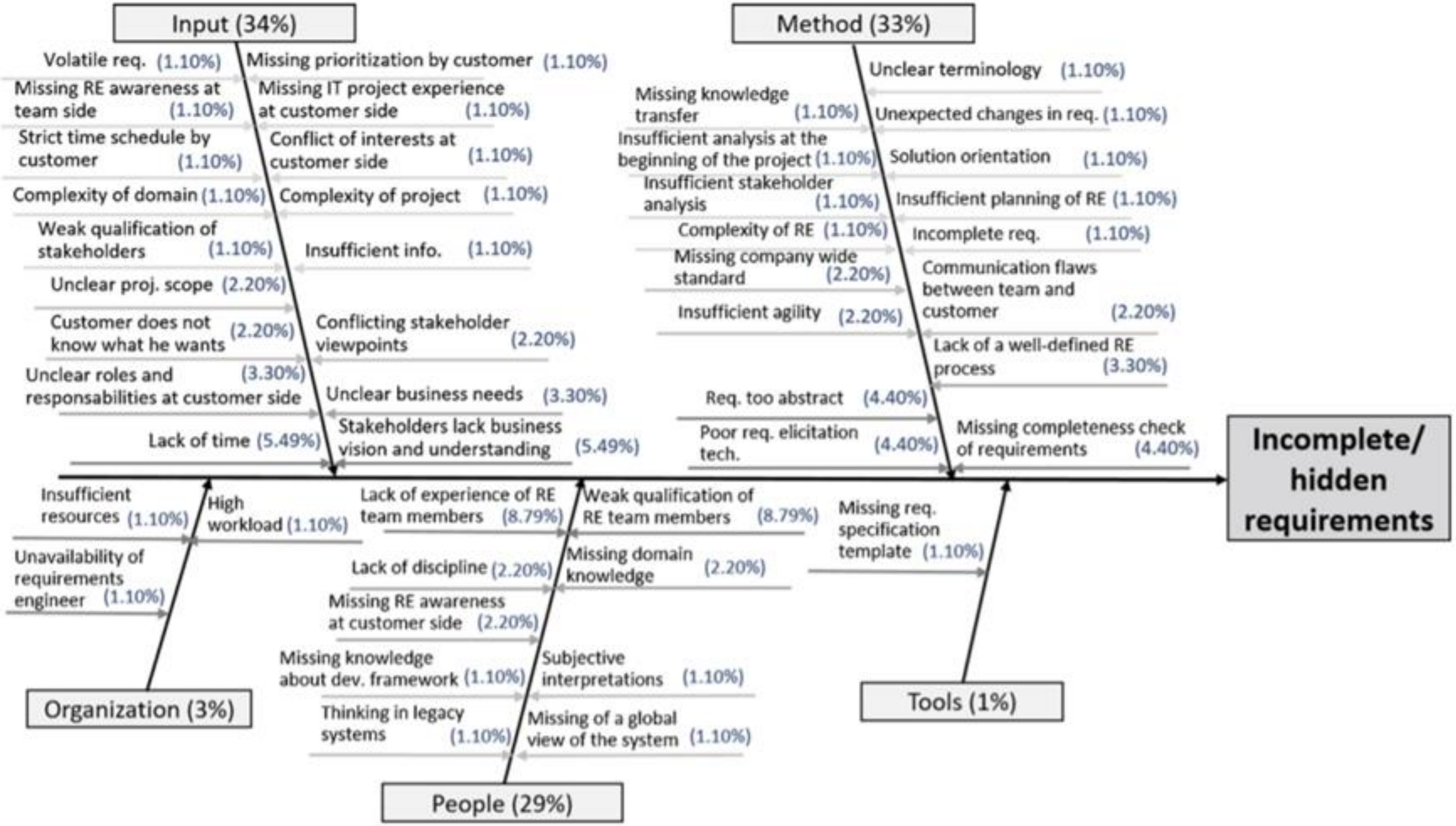

EXAMPLE

Qualitative Analysis: Example



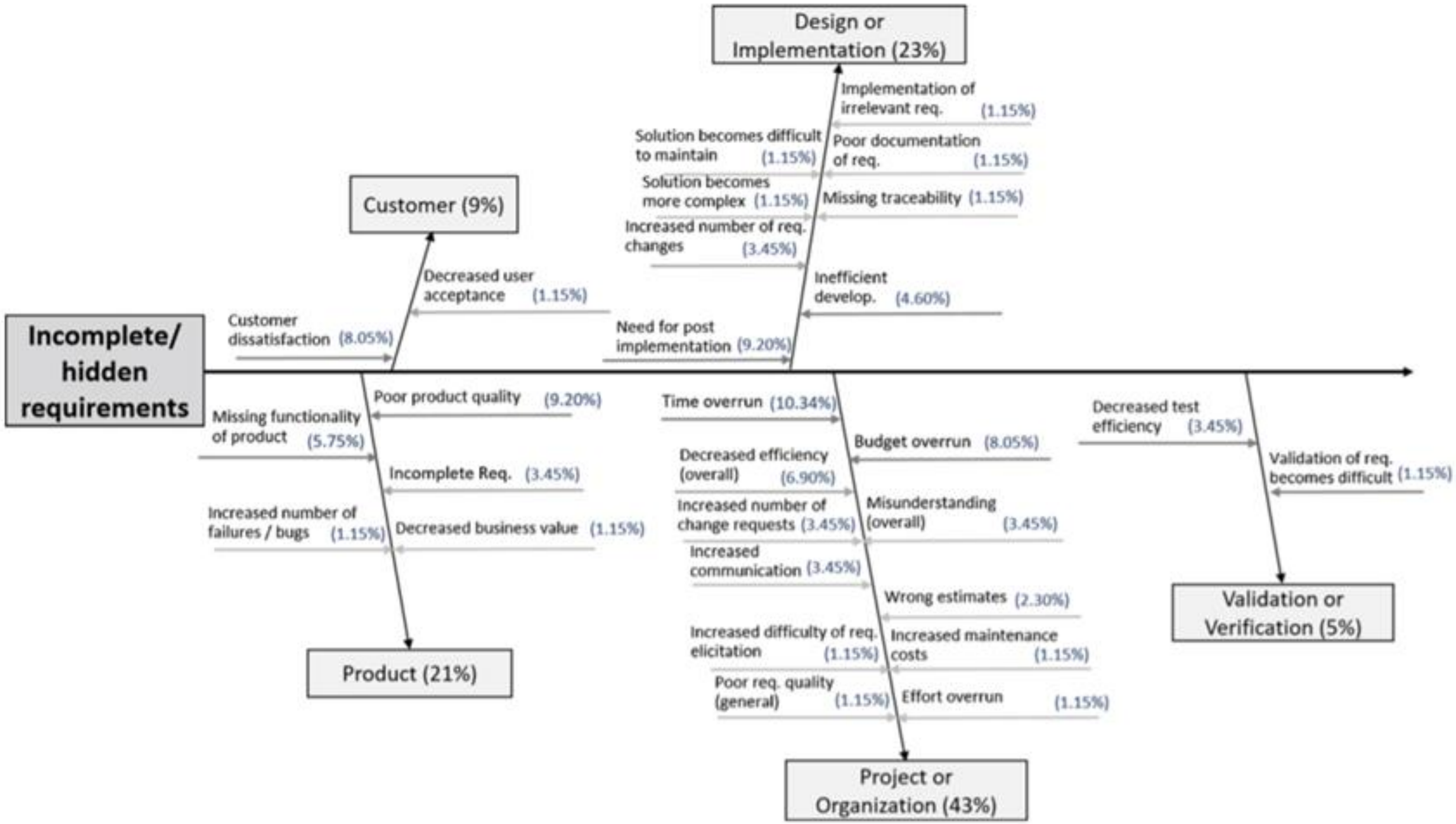
Fernández, D. M.; Wagner, S.; Kalinowski, M.; Felderer, M.; Mafra, P.; Vetro, A.; Conte, T.; Christiansson, M.; Greer, D.; Lassenius, C.; Männistö, T.; Nayabi, M.; Oivo, M.; Penzenstadler, B.; Pfahl, D.; Prikladnicki, R.; Ruhe, G.; Schekelmann, A.; Sen, S.; Spínola, R. O.; Tuzcu, A.; de la Vara, J. L.; and Wieringa, R. **Naming the pain in requirements engineering - Contemporary problems, causes, and effects in practice**. Empirical Software Engineering, 22(5): 2298-2338. 2017.

Qualitative Analysis: Example



Fernández, D. M.; Wagner, S.; Kalinowski, M.; Felderer, M.; Mafra, P.; Vetro, A.; Conte, T.; Christiansson, M.; Greer, D.; Lassenius, C.; Männistö, T.; Nayabi, M.; Oivo, M.; Penzenstadler, B.; Pfahl, D.; Prikladnicki, R.; Ruhe, G.; Schekelmann, A.; Sen, S.; Spínola, R. O.; Tuzcu, A.; de la Vara, J. L.; and Wieringa, R. **Naming the pain in requirements engineering - Contemporary problems, causes, and effects in practice**. Empirical Software Engineering, 22(5): 2298-2338. 2017.

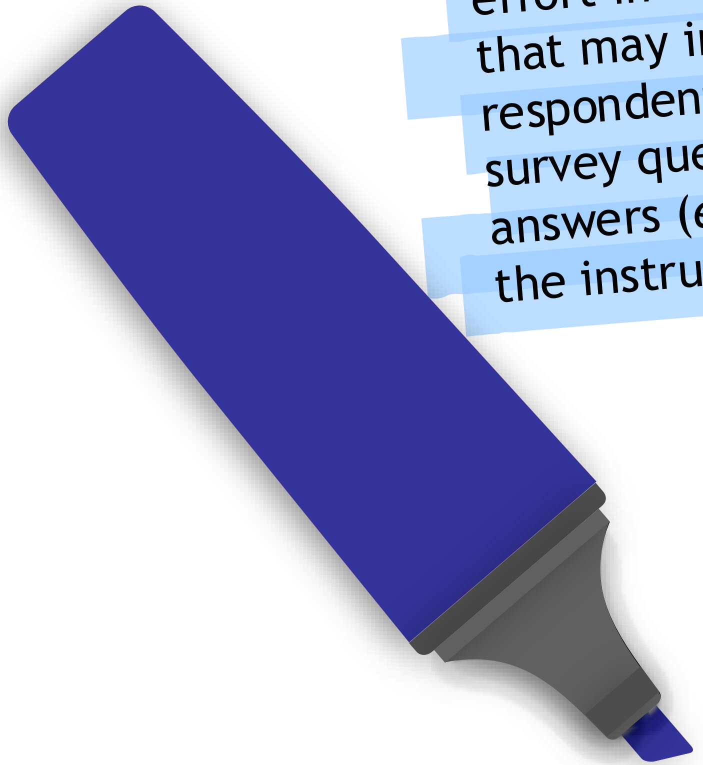
Qualitative Analysis: Example



Fernández, D. M.; Wagner, S.; Kalinowski, M.; Felderer, M.; Mafra, P.; Vetro, A.; Conte, T.; Christiansson, M.; Greer, D.; Lassenius, C.; Männistö, T.; Nayabi, M.; Oivo, M.; Penzenstadler, B.; Pfahl, D.; Prikladnicki, R.; Ruhe, G.; Schekelmann, A.; Sen, S.; Spínola, R. O.; Tuzcu, A.; de la Vara, J. L.; and Wieringa, R. **Naming the pain in requirements engineering - Contemporary problems, causes, and effects in practice**. Empirical Software Engineering, 22(5): 2298-2338. 2017.

DEPARTAMENTO DE INFORMÁTICA PUC-RIO

Key Takeaways *(Wagner et al., 2020)*:



A

When preparing your survey, invest effort in avoiding confounding factors that may interfere in having respondents focusing mainly on the survey question when providing their answers (*e.g.*, language issues). Assess the instrument validity.

B

Applying coding and analysis techniques from Grounded Theory can help to understand qualitative data gathered through open questions.

C

When reporting the qualitative analysis of your survey, explicitly state your research method, providing details on eventual deviations.

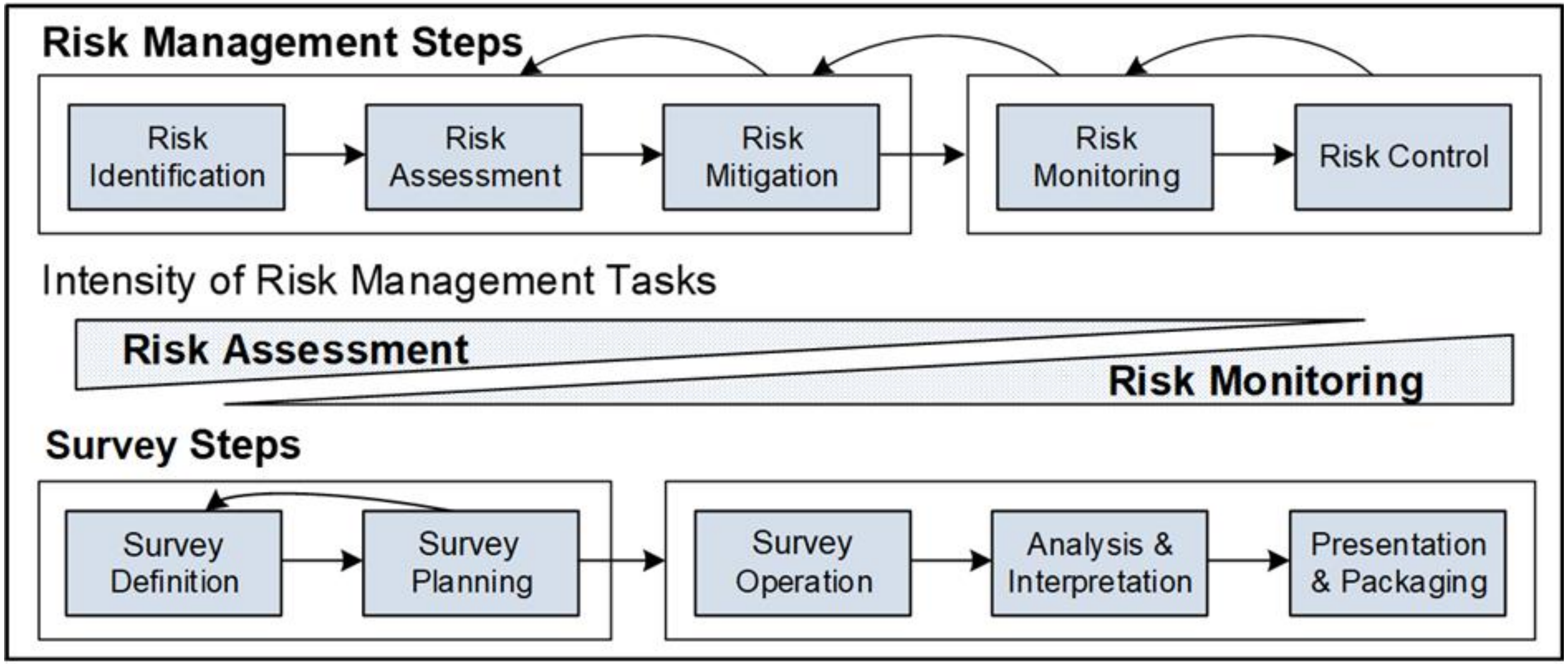
D

To avoid researcher bias and improve the reliability of the results, qualitative analyses should be conducted in teams and make use of independent validations. Also, ideally the raw and analyzed data should be open to enable other researchers to replicate the analysis procedures.

THREATS TO VALIDITY AND RELIABILITY

Survey Risk Management

- Validity is a property of inferences and every study faces **Threats to Validity** *(Biffl et al., 2014)*



Biffl, S., Kalinowski, M., Ekaputra, F., Neto, A.A., Conte, T. and Winkler, D., 2014, September. **Towards a semantic knowledge base on threats to validity and control actions in controlled experiments.** In *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (pp. 1-4).

Validity Assessment

In **psychometrics**, validity concerns "the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" *(AERA et al., 2014)*

Rust (2009) summarized six facets of validity in the context of psychometric tests:

- FACE VALIDITY
- CONTENT VALIDITY (*aka* criterion validity in this context)
- PREDICTIVE VALIDITY
- CONCURRENT VALIDITY
- CONSTRUCT VALIDITY
- DIFFERENTIAL VALIDITY

In **software engineering** we typically aim at assessing whether it is possible to safely conclude that a survey measures what it is supposed to

The following validity types are discussed in this context *(Kitchenham and Pfleeger, 2008 apud Linaker et al., 2015)*:

FACE VALIDITY

Typically involves a lightweight review of the questionnaire by randomly chosen respondents

CONTENT VALIDITY

Typically involves having a (focus) group of reviewers evaluating the questionnaire. The group should include subject matter experts and example respondents from the target population

CRITERION VALIDITY

Refers to how the questionnaire can separate between respondents that belong to different groups. An existing classification and mapping of the different groups in the target population must be in place

CONSTRUCT VALIDITY

validity is how well the question actually measures the construct it was intended to by the designer

Reliability (*aka* External Validity and Generalizability):

TEST-RETEST RELIABILITY

- ✓ The same subject responds to the same survey two times, and it is measured whether the subject gives the same answers each time
- ✓ Kitchenham and Peeger (2008) state that if the correlation between both of the answers is greater than 0.7 the test-retest reliability can be considered good

INTER-OBSERVER RELIABILITY

- ✓ Assesses observer interview bias in not self-administered surveys
- ✓ Assesses observer analysis bias (e.g., when interpreting and decoding open ended questions)
- ✓ Typically addressed by having two or more observers involved in the interview and analysis process

PHRASING / REORDER EFFECT RELIABILITY

- ✓ Testing whether the phrasing or reordering of questions has any effect on the answers by a respondent (assesses instrument bias on the respondent)

STATISTICAL GENERALIZABILITY

- ✓ If conclusions are to be drawn on the whole population, not just on the sample, the reliability needs to be proven and established


EXAMPLE

Threats to Validity and Reliability (Example)

Threats	Treatment
Face Validity – Bad instrumentation	Revision and evaluation of the questionnaire about the format and formulation of the questions. Questions objectively focused on the 3PDF. Running a pilot study.
Content Validity – Inadequate explanation of the constructs	Revision and evaluation of the questionnaire about the format and formulation of the questions. Running a pilot study. Providing a brief explanation on the 3PDF and a link with further details.
Criterion Validity – Not surveying the target population.	We identified SE SLR update authors following an explicitly documented and carefully conducted procedure (cf. Section 3).
Construct Validity – Inadequate measurement procedures and unreliable results.	We only used frequency counting, which can be safely applied to discrete survey questions concerning the relevance of the 3PDF questions and the agreement with the 3PDF decision drivers. Also, we triangulated the answers with the provided explanations.
Reliability – Lack of statistical conclusion validity	This threat strongly depends on the sample size. Unfortunately, while contacting twice the SE SLR update authors we were aware of, our final sample size was still limited. Hence, we focused our results on qualitative analyses and did not make any further claims on conclusion validity.

Mendes, E., Wohlin, C., Felizardo, K. and Kalinowski, M., 2020. **When to update systematic literature reviews in software engineering.** Journal of Systems and Software, 167, p.110607.

HANDS-ON EXERCISE

Developing a Short Survey

- Objective: Develop a survey relating **Happiness** with **Perceived Productivity** for Software Engineering Professionals.
    1. Individual Preparation (15min)
        - Take notes on:
            - » **Demographic information**
            - » **Constructs**
            - » **Hypotheses (Propositions)**

    2. Team Discussion (15min)
        - Use the Interactive Miro Board to discuss relevant demographic information, constructs, and hypotheses with the remaining participants.

- Material:
    - Scale of Positive and Negative Experience (SPANE) for Hapiness (http://labs.psychology.illinois.edu/~ediener/SPANE.html)
    - WHO's Health and Work Performance Questionnaire (HPQ) for Perceived Performance (https://www.hcp.med.harvard.edu/hpq/info.php)
    - Stack Overflow Survey for Demographics (https://insights.stackoverflow.com/survey/2021)

SPANE

Scale of Positive and Negative Experience (SPANE)

© Copyright by Ed Diener and Robert Biswas-Diener, January 2009.

Please think about what you have been doing and experiencing during the past four weeks. Then report how much you experienced each of the following feelings, using the scale below. For each item, select a number from 1 to 5, and indicate that number on your response sheet.

1. Very Rarely or Never
2. Rarely
3. Sometimes
4. Often
5. Very Often or Always

Positive
Negative
Good
Bad
Pleasant
Unpleasant
Happy
Sad
Afraid
Joyful
Angry
Contented

Scoring:
The measure can be used to derive an overall affect balance score, but can also be divided into positive and negative feelings scales.

Positive Feelings (SPANE-P): Add the scores, varying from 1 to 5, for the six items: positive, good, pleasant, happy, joyful, and contented. The score can vary from 6 (lowest possible) to 30 (highest positive feelings score).

Negative Feelings (SPANE-N): Add the scores, varying from 1 to 5, for the six items: negative, bad, unpleasant, sad, afraid, and angry. The score can vary from 6 (lowest possible) to 30 (highest negative feelings score).

Affect Balance (SPANE-B): The negative feelings score is subtracted from the positive feelings score, and the resultant difference score can vary from -24 (unhappiest possible) to 24 (highest affect balance possible). A respondent with a very high score of 24 reports that she or he rarely or never experiences any of the negative feelings, and very often or always has all of the positive feelings.

HPQ

B12.The next questions are about the time you spent during your hours at work in the past 4 weeks (28 days). Select the one response for each question that comes closest to your experience.

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
B12a. How often was your performance higher than most workers on your job?	O	O	O	O	O
B12b. How often was your performance lower than most workers on your job?	O	O	O	O	O
B12c. How often did you do no work at times when you were supposed to be working?					
B12d. How often did you find yourself not working as carefully as you should?					
B12e. How often was the quality of your work lower than it should have been?					
B12f. How often did you not concentrate enough on your work?					
B12g. How often did health problems limit the kind or amount of work you could do?					

B13. On a scale from 0 to 10 where 0 is the worst job performance anyone could have at your job and 10 is the performance of a top worker, how would you rate the usual performance of most workers in a job similar to yours?

Worst Performance 0	1	2	3	4	5	6	7	8	9	Top Performance 10
O	O	O	O	O	O	O	O	O	O	O

B14. Using the same 0-to-10 scale, how would you rate your usual job performance over the past year or two?

Worst Performance 0	1	2	3	4	5	6	7	8	9	Top Performance 10
O	O	O	O	O	O	O	O	O	O	O

B15. Using the same 0-to-10 scale, how would you rate your overall job performance on the days you worked during the past 4 weeks (28 days)?

Worst Performance 0	1	2	3	4	5	6	7	8	9	Top Performance 10
O	O	O	O	O	O	O	O	O	O	O

B16. How would you compare your overall job performance on the days you worked during the past 4 weeks (28 days) with the performance of most other workers who have a similar type of job? (Select only one.)

- O You were a lot better than other workers
- O You were somewhat better than other workers
- O You were a little better than other workers
- O You were about average
- O You were a little worse than other workers
- O You were somewhat worse than other workers
- O You were a lot worse than other workers

Exercise Discussion

- https://miro.com/app/board/uXjVPZDlxGo=/

References

- AERA, APA, NCME (2014) **Standards for educational and psychological testing**. American Educational Research Association, Washington
- Basili, V.R. and Rombach, H.D., 1988. **The TAME project: Towards improvement-oriented software environments**. IEEE Transactions on software engineering, 14(6), pp.758-773.
- Batista, M., Magdaleno, A. and Kalinowski, M., 2017, May. **A Survey on the use of Social BPM in Practice in Brazilian Organizations**. In Anais do XIII Simpósio Brasileiro de Sistemas de Informação (pp. 436-443). SBC.
- Biffl, S., Kalinowski, M., Ekaputra, F., Neto, A.A., Conte, T. and Winkler, D., 2014, September. **Towards a semantic knowledge base on threats to validity and control actions in controlled experiments.** In Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (pp. 1-4).
- Binning JF (2016) **Construct**. https://www.britannica.com/science/construct
- Charmaz, K. 2014. **Constructing Grounded Theory**. Sage, 2nd Ed.
- Corbin, J.M. and Strauss, A., 1990. **Grounded theory research: Procedures, canons, and evaluative criteria.** Qualitative sociology, 13(1), pp.3-21.
- Fernández, D. M.; Wagner, S.; Kalinowski, M.; Felderer, M.; Mafra, P.; Vetro, A.; Conte, T.; Christiansson, M.; Greer, D.; Lassenius, C.; Männistö, T.; Nayabi, M.; Oivo, M.; Penzenstadler, B.; Pfahl, D.; Prikladnicki, R.; Ruhe, G.; Schekelmann, A.; Sen, S.; Spínola, R. O.; Tuzcu, A.; de la Vara, J. L.; and Wieringa, R. **Naming the pain in requirements engineering - Contemporary problems, causes, and effects in practice**. Empirical Software Engineering, 22(5): 2298-2338. 2017.
- Fowler Jr, F.J., 2013. **Survey research methods.** Sage publications.
- Genero, M., Cruz-Lemus, J.A. y Piattini, M., 2014. **Métodos de investigación en ingeniería del software.** Ed. RA-MA.
- Glaser, B.G. 1992. **Basics of Grounded Theory Analysis: Emergence vs Forcing**. Sociology Press.
- Graziotin, D. and Fagerholm, F., 2019. **Happiness and the productivity of software engineers**. In Rethinking Productivity in Software Engineering (pp. 109-124). Apress, Berkeley, CA.
- Hannay, J.E., Sjoberg, D.I. and Dyba, T., 2007. **A systematic review of theory use in software engineering experiments**. IEEE transactions on Software Engineering, 33(2), pp.87-107.

References

- Kalinowski, M., Weber, K., Santos, G., Franco, N., Duarte, V. and Travassos, G., 2015. **Software Process Improvement Results in Brazil Based on the MPS-SW Model**. Software Quality Professional, 17(4): 15-28.
- Kalinowski, M., Weber, K., Franco, N., Barroso, E., Duarte, V., Zanetti, D. and Santos, G., 2014, September. **Results of 10 years of software process improvement in Brazil based on the MPS-SW model**. In 2014 9th International Conference on the Quality of Information and Communications Technology (pp. 28-37).
- Kalinowski, M., Weber, K.C. and Travassos, G.H., 2008, October. **iMPS: an experimentation based investigation of a nationwide software development reference model.** In Proceedings of the Second ACM-IEEE international symposium on Empirical Software Engineering and Measurement (ESEM).
- Kitchenham, B.A. and Pfleeger, S.L., 2008. **Personal opinion surveys**. In Guide to advanced empirical software engineering (pp. 63-92). Springer, London.
- Kuhrmann, M., Tell, P., Hebig, R. et al. **What Makes Agile Software Development Agile?** Submitted to Transactions on Software Engineering (2021).
- Linåker, J., Sulaman, S.M., Maiani de Mello, R. and Höst, M., 2015. **Guidelines for conducting surveys in software engineering.** Technical Report.
- Mendes, E., Wohlin, C., Felizardo, K. and Kalinowski, M., 2020. **When to update systematic literature reviews in software engineering**. Journal of Systems and Software, 167, p.110607.
- Mendoza, I., Kalinowski, M., Souza, U. and Felderer, M., 2019, January. **Relating verification and validation methods to software product quality characteristics: results of an expert survey**. In Proc. of the Software Quality Days Conference (SWQD) (pp. 33-44). Springer, Cham.
- de Mello RM, da Silva PC, Travassos GH (2015) **Investigating probabilistic sampling approaches for large-scale surveys in software engineering**. Journal of Software Engineering Research and Development, 3(1):8.
- Ralph, P., Baltes, S., Adisaputri, G., Torkar, R., Kovalenko, V., Kalinowski, M., Novielli, N., Yoo, S., Devroey, X., Tan, X., Zhou, M., Turhan, B., Hoda, R., Hata, H., Robles, G., Fard, A. M., and Alkadhi, R, **Pandemic Programming How COVID-19 affects software developers and how their organizations can help**. Empirical Software Engineering (2020), 25: 4927-4961. 2020.

References

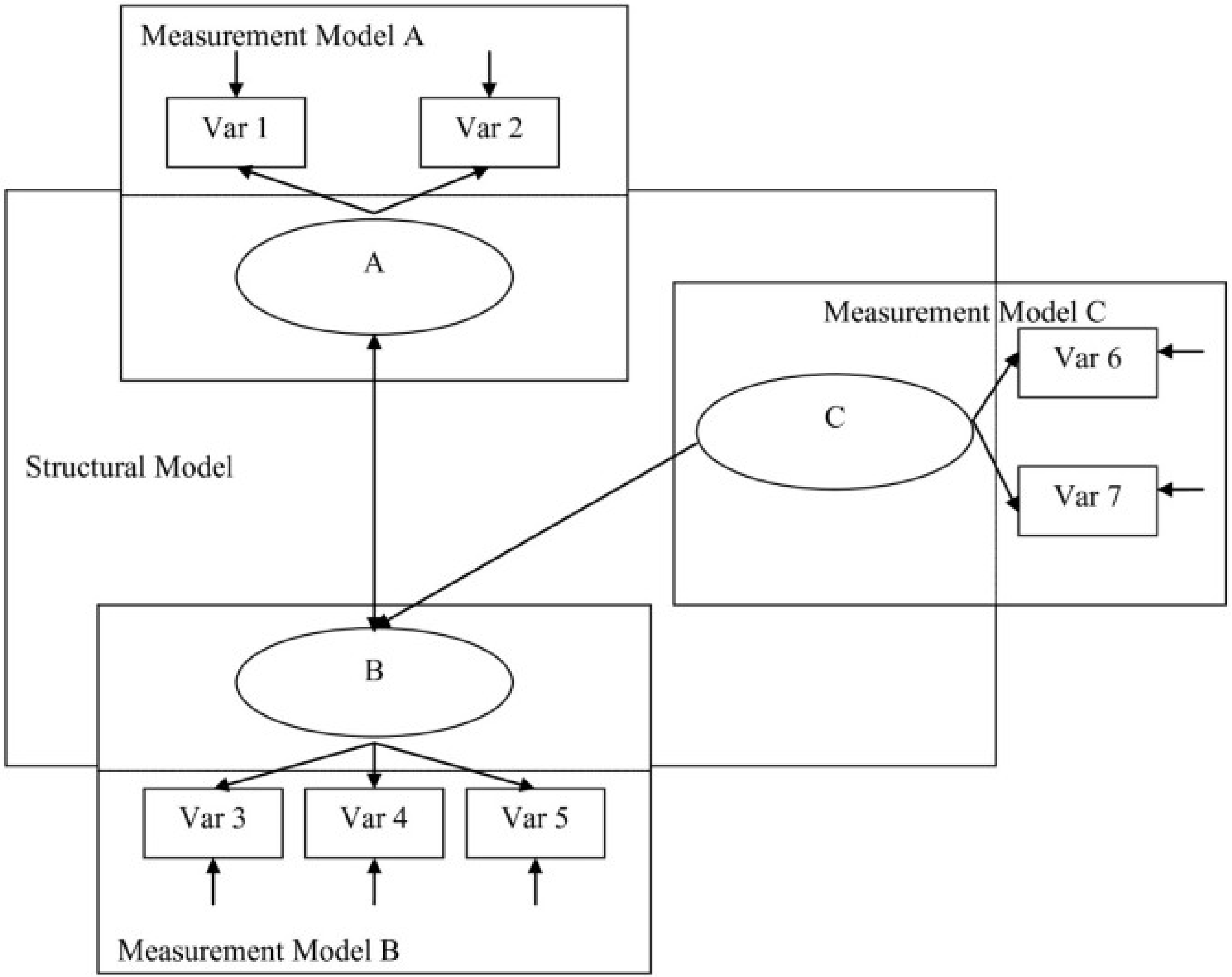
- Robson, C., (2002) **Real World Research - A Resource for Social Scientists and Practitioner-Researchers**, 2nd ed. Malden: Blackwell Publishing.
- Rust J (2009) **Modern psychometrics: the science of psychological assessment**. Routledge, Hove, East Sussex New York
- Sjøberg, D.I., Dybå, T., Anda, B.C. and Hannay, J.E., 2008. **Building theories in software engineering**. In Guide to advanced empirical software engineering (pp. 312-336).
- Stol, K.J., Ralph, P. and Fitzgerald, B., 2016, May. **Grounded theory in software engineering research: a critical review and guidelines**. In Proceedings of the 38th International Conference on Software Engineering (pp. 120-131).
- Torkar, R., Feldt, R. and Furia, C.A., 2020. **Bayesian Data Analysis in Empirical Software Engineering: The Case of Missing Data.** In Contemporary Empirical Methods in Software Engineering (pp. 289-324). Springer, Cham.
- Travassos, G.H. and Kalinowski, M., 2014. **iMPS 2013: Evidence on performance of organizations that adopted the MPS-SW**. Campinas, Brazil: Softex.
- Wagner, S., Mendez, D., Felderer, M., Graziotin, D. and Kalinowski, M., 2020. **Challenges in survey research.** In: Contemporary Empirical Methods in Software Engineering (pp. 93-125). Springer, Cham.
- Wagner, S., Fernández, D. M., Felderer, M., Vetro, A., Kalinowski, M., Wieringa, R., Pfahl, D., Conte, T., Christiansson, M., Greer, D., Lassenius, C., Männistö, T., Nayebi, M., Oivo, M., Penzenstadler, B., Prikladnicki, R., Ruhe, G., Schekelmann, A., Sen, S., Spínola, R.O., Tuzcu, A., de la Vara, J. L., and Winkler, D, **Status Quo in Requirements Engineering: A Theory and a Global Family of Surveys.** ACM Transactions on Software Engineering and Methodology, 28(2): 9:1-9:48. 2019.
- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B. and Wesslén, A., 2012. **Experimentation in software engineering**. Springer Science & Business Media.

SURVEY RESEARCH METHODS IN SOFTWARE ENGINEERING

Prof. Marcos Kalinowski

kalinowski@inf.puc-rio.br
http://www.inf.puc-rio.br/~kalinowski
@prof_kalinowski

DEPARTAMENTO DE INFORMÁTICA
PUC-RIO


Measurement Model A
Var 1
Var 2
A
Measurement Model C
Var 6
C
Structural Model
Var 7
B
Var 3
Var 4
Var 5
Measurement Model B