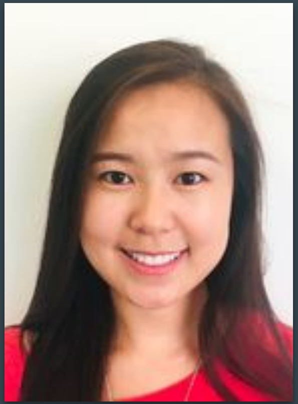
Verbal Data in Software Engineering: Challenges and Opportunities



Westley **Weimer**, University of Michigan

Yu **Huang**, Vanderbilt University

March 18, 2022 @ HUMAN 2022

Thesis

With appropriate care in data gathering and analysis, verbal data *can* provide impactful insights in software engineering research.

We believe verbal data to be particularly useful for overturning conventional wisdom and discovering unknown themes.

Outline (45 + 10)

- Verbal Data
    - Definitions, Metrics
    - Combining Verbal and Nonverbal Data
- Three Case Studies
    - Retrospective Recollections & Medical Imaging
    - Semi-Structured Surprises & Open Source for Social Good
    - Vulnerable Surveys & Climate Interviews
- Useful Techniques
        - Grounded Theory, Inductive Thematic Analysis, Inter-Rater Reliability
- Conclusion

Verbal Data

Verbally-acquired data

Information that is gathered via speech, think-aloud protocol, oral retrospection, formal or informal interviews

Classic Example: The "Sillito et al." Questions

Published in FSE '06, cited over 350 times

Questions Programmers Ask During Software Evolution Tasks

Jonathan Sillito, Gail C. Murphy and Kris De Volder
Department of Computer Science
University of British Columbia
Vancouver, B.C. Canada
{sillito,murphy,kdvolder}@cs.ubc.ca

them. Participants in the second study (E1…E16) were observed working on code with which they had experience. In both studies

During each session an audio recording was made of discussion between the pair of participants, a video of the screen was captured,

To structure our data collection and the analysis of our results, we have used a *grounded theory* approach which has been described as an emergent process intended to support the production of a theory that "fits" or "works" to explain a situation of interest [5, 19]. In

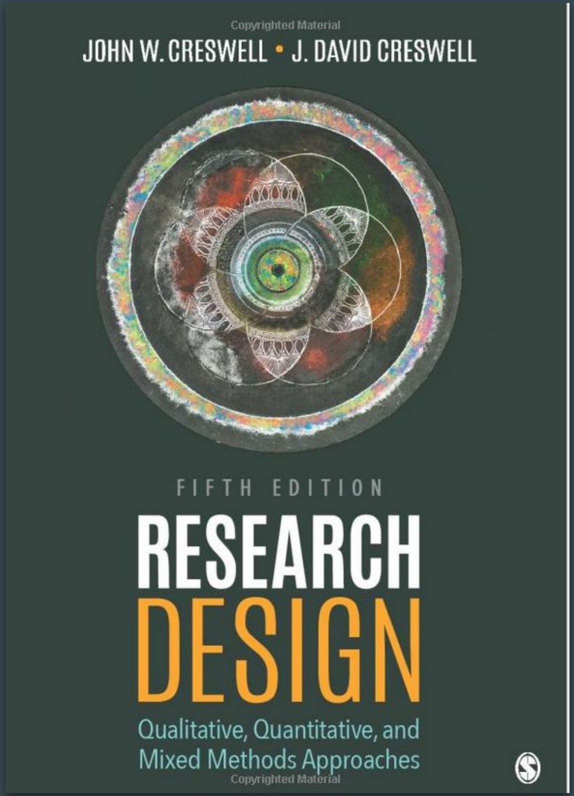
about the source code on which we observed them working. We report on 44 kinds of questions we observed our participants asking. These questions are generalized versions of the specific ques-

Results are useful directly (a structured answer to a fundamental question) and also as artifacts (re-used by later projects as indicative developer queries)

[Sillito, Murph, De Volder. Questions programmers ask during software evolution tasks. FSE 2006.]

Verbal Data Metrics

- Establishing validity in qualitative research
  - Using multiple validity procedures
    - Member checking
    - Clarify bias
    - Spend prolonged time in the field
  - Using qualitative reliability
    - Document your procedures (scripts, codebook, etc.)
    - No drift in the definition of codes
    - Cross-check codes developed by different researchers



[Chi. Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. J. Learning Sciences 1997.]

Combining Verbal and Nonverbal Data

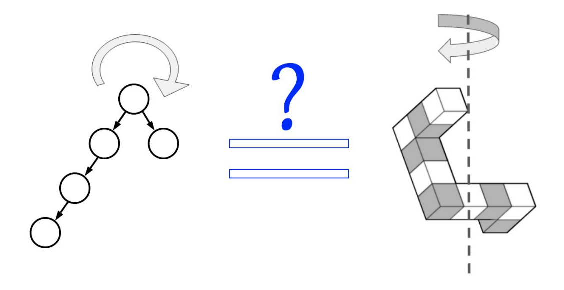
- Strength of verbal data
  - Richess and holism
  - Discovery
    - New ideas, hypothesis
- Weakness of verbal data
  - Hard to evaluate the analysis (i.e., no "equations")
  - Human biases
- Combining verbal and nonverbal data makes a strong and interesting case
  - Supplement, validate, or illuminate each other
  - Contrast: surprising knowledge!

Three Case Studies

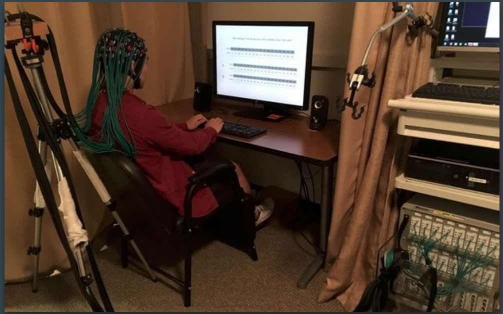
- Retrospective Recollections & Medical Imaging
    - Including aspects of “unreliable self-reporting” and “deception”
- Semi-Structured Surprises & Open Source for Social Good
    - Including aspects of “starting with no knowledge” and “surprised by your participants”
- Vulnerable Surveys & Climate Interviews
    - Including aspects of “protecting vulnerable populations”

Case Study 1a - Data Structures & Cognition

- How do human brains represent data structures? Is it more like text or more like 3D objects (mental rotation)?
- Nonverbal Data: Medical Imaging
  - fMRI
  - fNIRS
- Verbal Data: Interviews



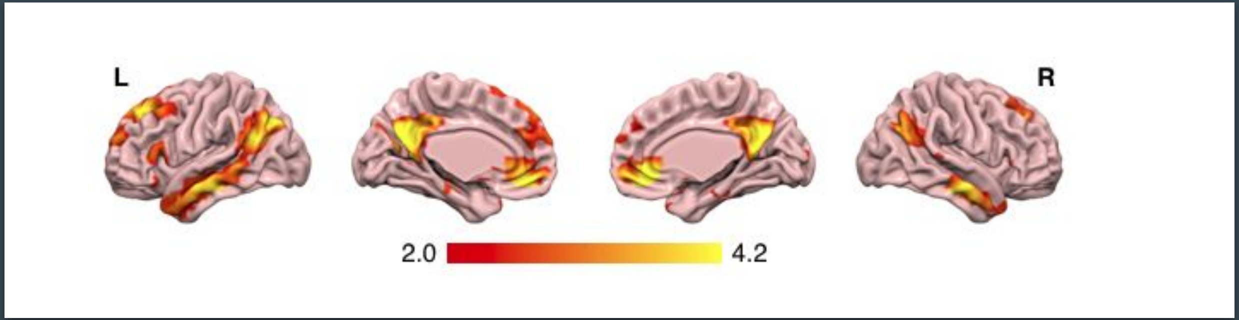
fMRI



fNIRS

Data Structures & Medical Imaging

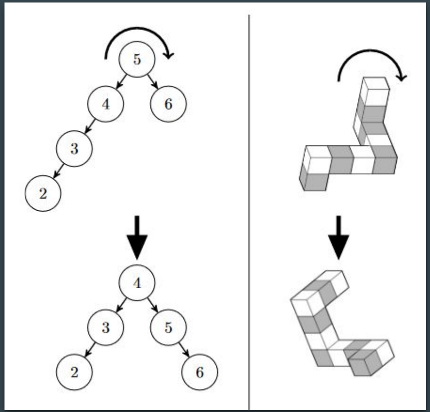
- What do we learn from nonverbal data (medical imaging)?
  - Data structure manipulations do use the **same** parts of the brain as rotating 3D objects
- Nonverbal data can be powerful!
  - You cannot just ask humans: “what do your brain patterns look like?”



Mental Rotation > Tree

Data Structures & Retrospective Reflection

- Verbal data
  - *“Do you think manipulating data structures and rotating 3D objects are similar tasks on the cognitive level?”*



Showing Prompts

Audio Recording

Transcribing

Qualitative Analysis

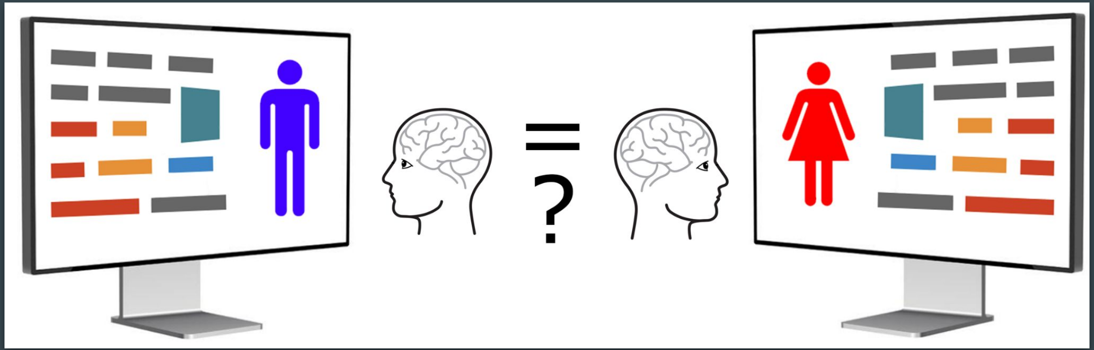
Data Structures & Combined Verbal and Nonverbal Data

- What do we learn from verbal data (audio / interviews)?
  - 70% of participants report **no similarity** between data structure manipulation and 3D object rotation
- Recall: What do we learn from nonverbal data (medical imaging)?
  - Data structure manipulations use the **same** parts of the brain as rotating 3D objects
- Why do we want to combine verbal and non-verbal information?
  - "Counterintuitive" knowledge from contrast (e.g., overturning conventional wisdom may inform or change how we teach or train going forward)
  - Learn the strength and weakness of both in different scenarios
    - Interpretability vs. Objectiveness

[Huang, Liu, Krueger, Santander, Hu, Leach, Weimer. Distilling Neural Representations of Data Structure Manipulation using fMRI and fNIRS. ICSE 2019.]

Case Study 1b - Bias in Code Review

- Is there bias related to gender and identities in code review? How do we characterize that bias?
  - Human vs. Machine (e.g., Automatic Program Repair, code synthesis, etc.)
  - Men vs. Women
- Can we just tell participants that we are investigating human bias on author information in code reviews?
    - Problem: social desirability bias!
    - Solution: deception in study design



Bias in Code Review & Deception

- Deception involves the justified use of false or misleading information
  - Sometimes it is necessary!
    - Hide the actual study goals, mitigate biases
  - IRB protocol approval: debriefing is required, cannot increase risk

Recruitment | **Deception** | **Code Reviews** | **Post Survey** | **Debriefing**

Deception:

(1) "We want to check how developers conduct code reviews"
(2) "We picked some real pull requests from software companies"
(3) "An author pic of a computer means it is generated by an algorithm"

Debriefing:

(1) "We are actually checking if author information affects your decision!"
(2) "All the author pictures are added purposely"
(3) "All pull requests are actually generated by humans"

Bias in Code Review & Free Response

- Could we ask the participants face to face?
    - "Do you think women and men write pull request differently?"
    - Writing down free responses or using solo recording is probably better!
- Self-reporting
    - "There is no difference between pull requests written by men and women"

→ **But there *is* a significant difference on your behavior! Both response time and final decisions are affected!**

    - "Machine generated code is worse on readability!"

→ **But all pull requests were written by humans! (We deceived you!)**

Combine "verbal" and nonverbal information, again!

[Huang, Leach, Sharafi, McKay, Santander, Weimer. Biases and Differences in Code Reviews using Medical Imaging and Eye-Tracking: Genders, Humans, and Machines. FSE 2020.]

Case Study 2 - Open Source Software for Social Good (OSS4SG)

- How can we characterize the OSS4SG community? How can we support them?
    - Technical Good vs. Social Good
- But we have barely any knowledge about the “social good” community!
    - (This is common when doing a first investigation into a phenomenon.)
    - What is “social good” in software?
    - We have only an ambiguous impression to start with …

REFUGE restrooms

Providing safe restroom access to transgender, intersex, and gender nonconforming individuals.

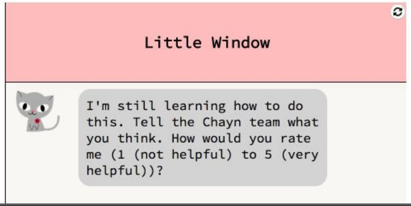
REFUGE is an effort to fill the void left by the now-defunct Safe2Pee website. It provides a free resource to trans* and queer individuals in need of gender neutral and other safe restrooms.

This project is open source. Feel free to contribute. We could use the help.

Little Window

> Little window is a clever little cat chatbot that directs women to the information they are looking for as quickly as possible. Think of it like google search on turbo for all of Chayn's resources and those provided by our friends too. We want to drastically reduce the time women take to search for information which in many cases can save lives.

Currently in active development build passing



Open Source for Social Good - Semi-Structured Interviews

- 21 one-hour-long interviews
  - Participants from all over the world, different cultural backgrounds
  - Online interviews (verbal data)
  - Audio recordings and transcriptions
  - Formal inductive thematic analysis
- Be prepared!
  - "Surprises" happen all the time – Precious!

ID	Gender	SG Exp	OSS Exp	Location of Contribution	Project Domains
P1	W	2	2	Mexico	Crypto, Security
P2	M	1	1	USA	Finance
P3	M	8	8	Germany	Education, Healthcare, Disaster Tracking
P4	W	1	1	UK	Charity, Domestic Violence
P5	M	1	3	India	Environment
P6	M	0.5	10	Turkey	COVID-19 Tracking
P7	M	0.5	0.5	India	Education, Environment
P8	M	4	5.5	Israel	Structurally-Safe Buildings
P9	M	8	8	Australia	Healthcare, Education
P10	W	2	2	India	Healthcare, Education
P11	W	0.5	0.5	India	Education
P12	M	2	2	USA	COVID-19 Tracking
P13	M	2	6	USA	Education, Non-profit Tools
P14	NB	8	8	Germany	Anti-Gentrification, Safe Restrooms
P15	M	10	10	Spain	eGovernment, Civil Participation
P16	M	0.5	0.5	India	Healthcare
P17	M	2	2	India	Education
P18	M	1.5	1.5	Romania	Local Administration
P19	M	0.5	1	India	Healthcare
P20	M	5	15	Canada	Management for Government and Charity
P21*	M	5	5	USA	Healthcare

Open Source for Social Good - Surprises

- "Positive" surprises
  - *"We all know that after a short time, no one is gonna use our software anymore. But we still spend a lot of time on it!"* – From a contributor in a COVID-19 related project

What is motivating you in this case?

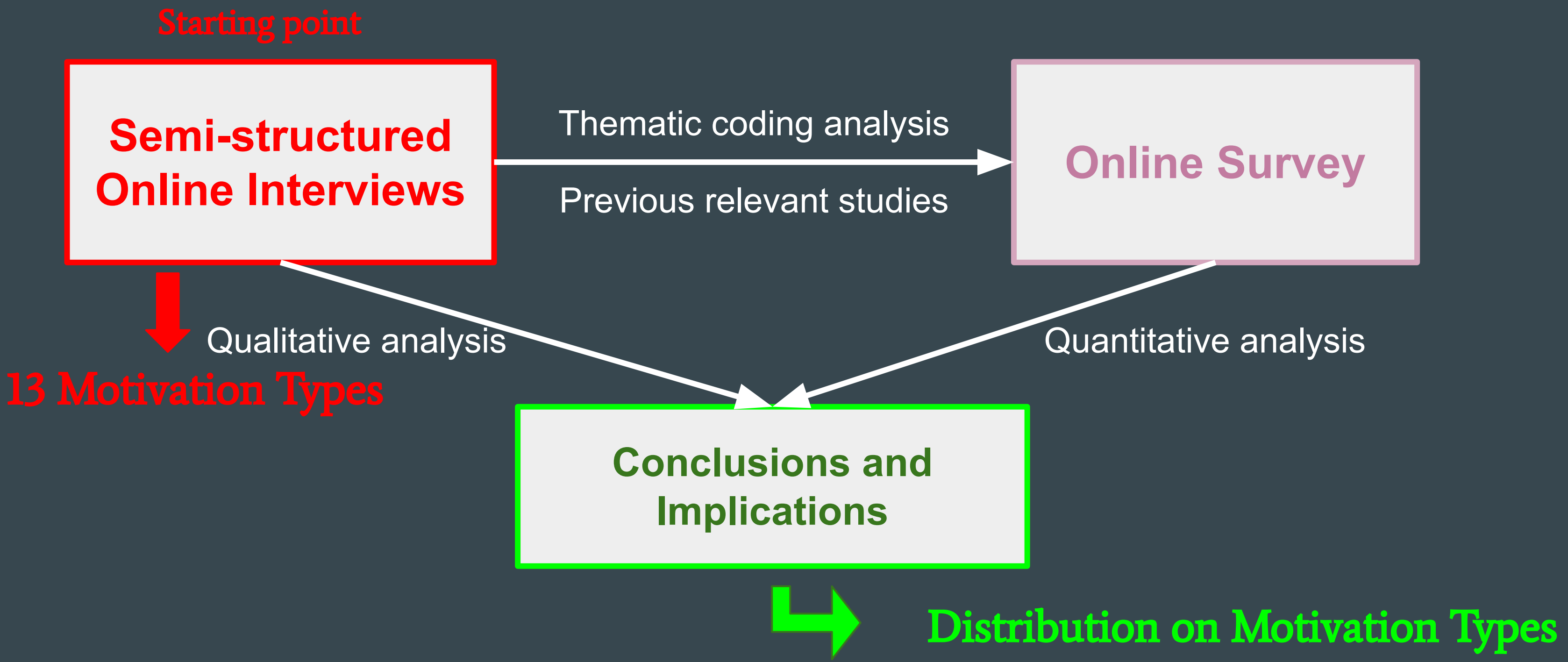
- "Negative" surprises
  - *"People here hate me after they know. Because they don't like it when you work for LGBTQ groups."* – From a contributor in a pro-LGBTQ project

How can we protect contributors like you?

- How can you respond to the surprises to get more information to answer your research questions?
  - Lesson learned: be prepared for rich information, requires sensitivity
  - Challenging but "feels so good"!

Open Source for Social Good - Results

- Power of verbal data: guidance



[Huang, Ford, Zimmermann. Leaving My Fingerprints: Motivations and Challenges of Contributing to OSS for Social Good. ICSE 2021.]

Case Study 3a - Cannabis Use in Software Engineering

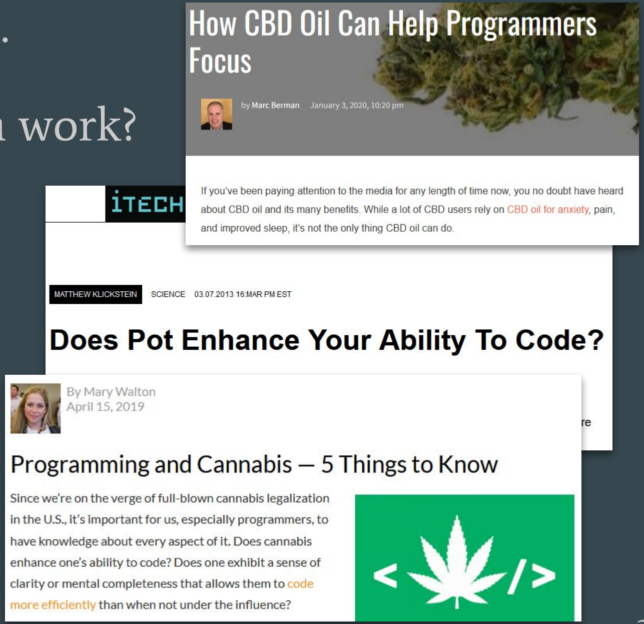
Cannabis is the world's most widely-used illicit substance, and its legality is changing rapidly. In the US, 17 states have legalized it for adult use despite it being a Schedule I drug at the federal level. It *is* used by some programmers.

- Folk wisdom: Does it help creativity? Hurt precision work?
- Employment: Are drug-test policies merited?


The FBI Says It Can't Find Hackers to Hire Because They All Smoke Pot

"I have to hire a great work force to compete with those cyber criminals and some of those kids want to smoke weed on the way to the interview."

By Max Cherney


How CBD Oil Can Help Programmers Focus

by Marc Berman January 3, 2020, 10:20 pm

If you've been paying attention to the media for any length of time now, you no doubt have heard about CBD oil and its many benefits. While a lot of CBD users rely on CBD oil for anxiety, pain, and improved sleep, it's not the only thing CBD oil can do.

MATTHEW KLICKSTEIN SCIENCE 03.07.2013 16:MAR PM EST

Does Pot Enhance Your Ability To Code?

By Mary Walton
April 15, 2019

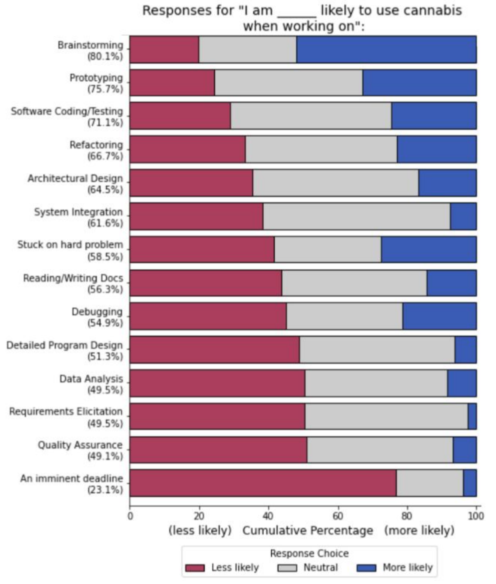
Programming and Cannabis — 5 Things to Know

Since we're on the verge of full-blown cannabis legalization in the U.S., it's important for us, especially programmers, to have knowledge about every aspect of it. Does cannabis enhance one's ability to code? Does one exhibit a sense of clarity or mental completeness that allows them to code more efficiently than when not under the influence?

Cannabis Use in Software Engineering - Informed Consent

Ethical research usually collects signed Informed Consent forms. Informally, for illegal activities, those forms would be signed confessions. IRBs may allow a Waiver of Informed Consent for vulnerable populations.

- We surveyed 800 developers (inc. 450 full-timers)
- Prevalence: 35% had programmed while using cannabis
- Perception: Devs expected managers to be negative on cannabis, but in practice managers were not ($p < 0.0001$)
- Freeform responses informed topics for subsequent semi-structured interviews.
- Recommendation: ask about follow-up contact.



[Endres, Boehnke, Weimer. Hashing It Out: A Survey of Programmers' Cannabis Usage, Perception, and Motivation. ICSE 2022.]

Case Study 3b - Climate Reporting in Computer Science

In previous years, the University of Michigan faced multiple allegations of faculty sexual misconduct. At the same time, world events (e.g., potential policy changes for international students) added stress and uncertainty.

- How can we proactively hear from students about climate (e.g., lab culture, personal experiences)?
- Uniform surveys were not a good fit (issues not known in advance, high variability in experiences, participant fear, etc.)

News

U of M computer science assistant professor resigns as sexual misconduct allegations come to light

Michigan Radio | By Nisa Khan
Published June 3, 2021 at 10:34 PM EDT

SEX, BOTS, AND AUDIOTAPE

At an AI startup with deep ties to the University of Michigan, employees allege a pattern of sexual harassment by the CEO

By Zoe Schiffer | @ZoeSchiffer | Feb 13, 2020, 2:45pm EST
Illustrations by Alex Castro

EDUCATION

ICE Threatens To Deport International Students If They Don't Attend In-Person Classes

July 7, 2020 · 3:57 PM ET
Heard on All Things Considered

MAX LARKIN (WBUR)

FROM wbur

Climate Reporting in Computer Science - Interviews

Challenges: fear of retaliation ("if I say anything negative, my advisor will …"), policy misunderstanding ("if I switch advisors, I will be deported"), cultural issues ("it is not appropriate to volunteer complaints or concerns, even if I have them")

Solution: 15-minute verbal interview ("check-in") with each graduate student (n > 300)

- Each interview is conducted by a staff member (non-faculty)
- Interviews are not recorded but notes are taken (and then coded)
- Questions are value-neutral: "how often do you meet with your advisor?", "describe an average work week", etc.
- Memorandum of Understanding (MOU) and explicit policy to destroy notes after analysis protect staff and students (beyond usual IRB protections!)

Climate Reporting in Computer Science - Summary

- Care must be taken when disaggregating results to retain anonymity
- Identified themes were supported both numerically and also via de-identified quotes (with indicative aspects highlighted)
- Identified surprises (e.g., no significant per-lab differences, job search assistance)
- Identified actionable positives (advisor communication, collaboration, work-life balance) and negatives (micromanagement, apathetic communication, lack of a second supporting faculty)

particular. The large number of Ph.D. student responses admits a focus on that subpopulation. The most-mentioned concerns were:

Work-life Balance	Sense of Community	Advisor Treatment	Lack of Collaboration
47	46	40	30
Switching Advisors	Faculty Access	Faculty Sexual Misconduct	Mental Health Resources
24	18	16	14

Indicative Ph.D. student quotes about mental health resources (emphasis ours):

- It's lacking in some places. **When you know where to look, it's there. But no one knows where to look.**
- The Engineering CAPS has been helpful, we can meet for a semester, but it's expected that it's temporary. It'd be nice if there was a CS specific resource. **[CAPS] Doesn't address student responsibilities and how to handle that.**
- During my graduate school experience, many of my friends share the feeling, that basically **it's very easy for us to feel depressed about the work we do.** Better emotional support, not just talking about research.

[Weimer. CSE Graduate Student Check-In Report. UM CSE Tech. Report 2021.]

Useful Techniques

In the last few slides, we summarize some useful techniques for getting started with verbal data

Grounded Theory in SE

IEEE TRANSACTIONS ON SOFTWARE ENGINEERING

Socio-Technical Grounded Theory for Software Engineering

Rashina Hoda

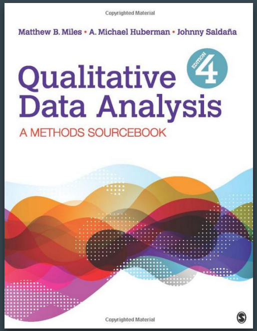
- Similar to socio-technical studies, qualitative research can have a lot of variance
  - How can we mitigate that variance?
- Grounded Theory is a systematic methodology for qualitative research for constructing hypotheses via inductive (not deductive) reasoning
  - Method
    - Empirical/evidence based
  - Outcome
    - Key patterns of the data
    - Relationships between patterns

“It is not in your mind; it is in your data.”

[Hoda. Socio-Technical Grounded Theory for Software Engineering. IEEE Trans. Software Engineering 2021.]

Grounded Theory in SE: Techniques

- Inductive Thematic Analysis
  - Thematic exploration
    - Codes and the relationships
    - E.g. Tesch's Eight-Step Coding Process
  - Evaluation metrics
    - Saturation
    - Agreement



Leaving My Fingerprints: Motivations and Challenges of Contributing to OSS for Social Good

Yu Huang
University of Michigan
Ann Arbor, MI
yhhy@umich.edu

Denae Ford
Microsoft Research
Redmond, WA USA
denae@microsoft.com

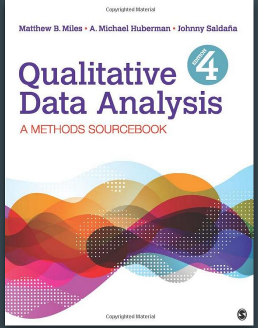
Thomas Zimmermann
Microsoft Research
Redmond, WA USA
tzimmer@microsoft.com

Codebook Example:

Category	Code	Description
motivation	motivation-helpuser	help end users
	motivation-helpdev	help developers
	motivation-longterm	how to keep yourself engaged in the project for a long time
	motivation-giveback	altruism
	motivation-impact	want to make impact
	motivation-better-programmer	want to look good in the community, improving skills, build up portofolio
	mitivation-hobby	I feel happy/fun, e.g., as a hobby.
	motivation-work	This is my job, or school projects, etc

Grounded Theory in SE: Techniques

- Inductive Thematic Analysis
  - Thematic exploration
    - Codes and the relationships
    - E.g. Tesch's Eight-Step Coding Process
  - Evaluation metrics
    - Saturation
    - Agreement
- Inter Rater Reliability (IRR) or Inter Rater Agreement (IRA)
  - Statistics as evidence
    - Cohen's kappa, Fleiss' kappa, etc.



Leaving My Fingerprints: Motivations and Challenges of Contributing to OSS for Social Good

Yu Huang
University of Michigan
Ann Arbor, MI
yhhy@umich.edu

Denae Ford
Microsoft Research
Redmond, WA USA
denae@microsoft.com

Thomas Zimmermann
Microsoft Research
Redmond, WA USA
tzimmer@microsoft.com

Conclusion

- Verbal data can require care in data gathering and analysis
  - Gathering: deception and value-neutral questions (avoid bias), waiver of informed consent / MOU (protection)
  - Analysis: inter-rater reliability (rigor), disaggregation (privacy), grounded theory (qualitative hypothesis discovery), inductive thematic analysis (coding)
- Verbal data can provide impactful insights in SE research
  - Overturning conventional wisdom ("data structures are not related to object rotation", "I am not biased by gender when I review code")
  - Discovering unknown themes ("I am worried about jobs, please spend more time preparing me", "people will stop using my software, but I am motivated anyway!")
- Research inherently involves risk and surprise: verbal data can be a powerful tool!