Measurement

Isaac Griffith

CS 6620 - Empirical Software Engineering
Department of Computer Science
Idaho State University

Inspiration

"You cannot control what you cannot measure" - DeMarco

Outline

- Introduction to measurement theory
- Scale types
- Threats to Validity

Introduction

Measurement provides a level of control necessary to conduct empirical studies and to manage projects, products, and processes

- **Measurement** - “Measurement is the process by which numbers or symbols are assigned to attributes of entities in the real world in such a way as to describe them according to clearly defined rules.”
- **Measure** - The number or symbol assigned to an entity by this relationship to characterize an attribute.
- **Metric** is a common word thrown around SE, and has two different meanings
  1. A term denoting the field of measurement in software engineering.
  2. An entity which is measured.

Measurement Theory

- A measure is a mapping from an entity's attribute to a value (usually numerical)
- This mapping allows the characterization and manipulation of the attributes in a formal way
- A measure must preserve the empirical observations of the attribute
- A **scale** is the representation of a mapping from an attribute to a measurement value.

Measurement Theory

- A measure of an attribute can be measure in different scales.
  - We can **transform** (or **rescale**) a measurement.
  - If the transform used preserves the relationship among objects it is an **admissible transform**
  - If the statements made about an object remain true after rescaling, the statements are called **meaningful** otherwise they are **meaningless**.
  - Example: if we have room A at 10C and room B at 20C, and make the statement room B is twice as warm as room A. If we then rescale to Fahrenheit we get 50F and 68F. The statement is no longer true and thus meaningless.

Measurement Validity

- To be valid:
  - a measure must not violate any necessary properties of the attribute it measures
  - it must be a proper mathematical characterization of the attribute
  - it must be both analytically and empirically valid
- Analytical Validity

Idaho State University
Computer Science

Scale Types

The four common scale types are:

- Nominal
- Ordinal
- Interval
- Ratio

ROAR

Scale Types

Research useful for

- Qualitative research is concerned with nominal and ordinal scales
- Quantitative research is concerned with interval and ratio scales

Properties

- Depending on which transform can be made on a scale, different scale types can be defined.
- Scales that belong to a scale type share the same properties.

Nominal Scale

- Least powerful of the scale types
- Maps attribute of the entity into a name or symbol
- Applicable transformations:
  - Any that preserve the that entities can be mapped one-to-one
- Examples:
  - Classification
  - Labeling
  - Defect typing

Ordinal Scale

- Ranks entities after an ordering criterion
- More powerful than Nominal scale
- Ordering Examples:
  - "greater than"
  - "better than"
  - "more complex"
- Applicable transformations:
  - Any which preserve the order of the entities: i.e. $M' = F(M)$
  - Where M' and M are different measures on the same attribute, and
  - F is a monotonic increasing function
- Examples:
  - Grades
  - Software complexity

Interval Scale

- More powerful than ordinal scale
- Used when the difference between two measures are meaningful, but not the value itself.
- Orders in the same way as an ordinal scale, but there is a notion of "relative distance" between two entities.
- Uncommon in Software Engineering
- Applicable transformations:
  - Where the measures are a linear combination of each other: $M' = \alpha M + \beta$
  - where M' and M are different measures on the same attribute.
- Examples:
  - temperature measured in Celsius or Fahrenheit

Ratio Scale

- More powerful than interval
- If there exists a meaningful zero value and the ratio between two measures is meaningful
- Possible transformations:
  - those that have the same zero and the scalar only differs by a factor: $M' = \alpha M$
  - where M' and M are different measures on the same attribute
- Examples:
  - length
  - temperature measured in Kelvin
  - duration of a development phase

Idaho State University
Computer Science

Scales Compared

Type	Meaning	Admissible Operations
Nominal Scale	Unordered classification of objects	=
Ordinal Scale	Ranking of objects into ordered categories	=, <, >, mode
Interval Scale	Differences between points on the scale are meaningful	=, <, >, difference, mean
Ratio Scale	Ratios between points on the scale are meaningful	=, <, >, difference, mode, mean, ratio
Absolute Scale	No units necessary - scale cannot be transformed	=, <, >, difference, mode, mean, ratio

ROAR

Measurement Types

- **Objective Measures**
  - No judgment in measurement
  - Value is only dependent on object
  - Can be measured several times without changing
- **Subjective Measures**
  - Depends on:
    - the object measured
    - the viewpoint of measurement
  - Values may change per measurement
  - Subject to potential bias
- **Direct Measures**
  - independent of other attributes/measures
  - Examples: LOC, $Defect_{test}$
- **Indirect Measures**
  - dependent on other attributes/measures
  - Examples: $Defect_{density} = \#Defects/LOC$, $Prod = LOC/effort$

Idaho State University Computer Science

SE Measurements

Classes of SE Objects for measurement:

- **Process** - The process describes which activities that are needed to produce the software
- **Product** - The products are the artifacts, deliverables or documents that results form a process activity
- **Resources** - Resources are the objects, such as personnel, hardware, or software, needed for a process activity

Type of attributes to measure:

- **Internal** - an attribute that can be measured purely in terms of the object
- **External** - an attribute that can only be measured with respect to how the object relates to other objects.

ROAR

Examples

Class	Example objects	Attribute type	Measure Example
Process	Testing	Internal External	Effort Cost
Product	Code	Internal External	Size Reliability
Resource	Personnel	Internal External	Age Productivity

Measurements in Practice

- Metrics are defined by the researcher and collected during the operation phase
- Metrics should be easy to collect
- Quality of collected measures is of utmost importance
- Understanding what is measured and its scale is important to determine what types of analysis are applicable
  - Know the distribution (uniform, normal, exponential, etc.)

Validity vs. Reliability

- Reliability: Does the study get consistent results?
- Validity: Does the study get true results?

Validity (positivist view)

- Construct Validity
  - Are we measuring the construct we intended to measure?
  - Did we translate these constructs correctly into observable measures?
  - Did the metrics we use have suitable discriminatory power?
- Internal Validity
  - Do the results really follow from the data?
  - Have we properly eliminated any confounding variables?
- External Validity
  - Are the findings generalizable beyond the immediate study?
  - Do the results support the claims of generalizability?
- Empirical Reliability
  - If the study was repeated, would we get the same results?
  - Did we eliminate all researcher biases?

Typical Problems

- Construct Validity
  - Using things that are easy to measure instead of the intended concept
  - Wrong scale; insufficient discriminatory power
- Internal Validity
  - Confounding variables: Familiarity and learning;
  - Unmeasured variables: time to complete task, quality of result, etc.
- External Validity
  - Task representativeness: toy problem?
  - Subject representativeness: students for professional developers!
- Theoretical Reliability
  - Researcher bias: subjects know what outcome you prefer

Construct Validity

- E.g. Hypothesis: “Inspection meetings are unnecessary”
  - Inspection -> Perspective-based reading of requirements docs
  - Meeting -> Inspectors gather together and report their findings
  - Unnecessary -> find fewer total # errors than inspectors working alone
- But:
  - What’s the theory here?
  - E.g. Fagin Inspections:
    - Purpose of inspection is process improvement (not bug fixing!)
    - Many intangible benefits: staff training, morale, knowledge transfer, standard setting, …

Construct Validity

- Are we measuring what we intend to measure?
  - Akin to the requirements problem: are we building the right system?
  - If we don't get this right, the rest doesn't matter
  - Helps if concepts in the theory have been precisely defined!
- Divide construct validity into three parts:
  - Intentional Validity- are we measuring precisely what we intend?
    - E.g. measuring "expertise" as "years of expertise"?
  - Representation Validity- do our measurements accurately operationalize the constructs?
    - E.g. is it okay to break "intelligence" down into verbal, spatial & numeric reasoning?
    - Face validity argument – "seems okay on the face of it"
    - Content validity argument – "measures demonstrated to cover the concept"
  - Observation Validity- how good are the measures by themselves?
    - E.g. the short form of a test correlates well with longer form

Observation Validity

- Predictive Validity
  - Observed measure predicts what it should predict and nothing else
    - E.g. check that college aptitude tests do predict college success
- Criterion Validity
  - Observed measure agrees with an independent standard
    - E.g., for college aptitude, GPA or successful first year
- Convergent Validity
  - Observed measure correlates with other observable measures for the same construct
    - I.e., our measure gives a new way of distinguishing a particular trait while correlating with similar measures
- Discriminant Validity
  - Observed measure distinguishes between two groups that differ on the trait in question
    - E.g. Measurement of code quality can distinguish “good” code from “bad”

Internal Validity

- Can we be sure our results really follow from the data?
  - Have we adequately ruled out rival hypotheses?
- Have we eliminated confounding variables?
  - Participant variables
  - Researcher variables
  - Stimulus, procedural and situational variables
  - Instrumentation
  - Nuisance variables

Internal Validity

- Confounding sources of internal invalidity
  - H: History
    - events happen during the study (e.g., study session was interrupted)
  - M: Maturation
    - older/wiser/better between treatments (or during study)
  - I: Instrumentation
    - change due to observation/measurement instruments
  - S: Selection
    - differing nature of participants
    - effects of choosing participants

External Validity

- Two issues:
  - Results will generalize beyond the specific situations studied
    - E.g. do results on students generalize to professionals?
  - Do the results support the claimsof generalizability?
    - E.g. if the effect size is small, will it be swamped/masked in other settings?
    - E.g. will other (unstudied) phenomena dominate?
- Two strategies:
  - Provide arguments in favor of generalizability
  - Replicate the finding in further studies
    - Literal replication - repeat study using the same design
    - Empirical induction - related studies test additional aspects of the theory
- Also: Ecological Validity
  - Does the study set-up approximate real-world conditions?
  - (can achieve external validity without this, but it's hard)

Reliability

- Could the study be repeated with the same results?
  - On the same subjects (not a replication!)
- Issues:
  - No mistakes were made in conducting the experiment
  - Steps taken in data collection and analysis were made explicit
  - No biases were introduced by the researchers
- Good practice
  - Carefully document all procedures used in the study
  - Prepare a “lab package” of all materials and procedures used
  - Conduct the study in such a way that an auditor could follow the documented procedures and arrive at the same results

Validity (Constructivist View)

- Repeatability is suspect:
  - Reality is “multiple and constructed”, same situation can never recur
  - Researcher objectivity is unattainable
  - E.g. successful replication depends on tacit knowledge
- Focus instead on “trustworthiness”:
  - Credibility of researchers and results
  - Transferability of findings
  - Dependability - results are robust across a range of situations
  - Confirmability
- Identify strategies to increase trustworthiness…

Strategies for constructivists

- Triangulation
  - Different sources of data used to confirm findings
- Member checking
  - Research participants confirm that results make sense from their perspective
- Rich, think descriptions
  - As much detail as possible on the setting and the data collected
- Clarify bias
  - Be honest about researcher's bias
  - Self-reflection when reporting findings
- Report discrepant information
  - Include data that contradicts findings as well as that which confirms
- Prolonged contact with participants
  - Spend long enough to ensure researcher really understands the situation being studied
- Peer debriefing
  - A colleague critically reviews the study and tests assumptions
- External Auditor
  - Independent expert reviews procedures and findings

Are there any questions?