Software Metrics
Threats to Validity

Presented at CAST 2012
July 17, 2012, San Jose, CA

Cem Kaner, J.D., Ph.D.

Nawwar Kabbani, M.Sc.

Florida Institute of Technology

These notes are partially based on research that was supported by NSF Grant CCLI-0717613 "Adaptation & Implementation of an Activity-Based Online or Hybrid Course in Software Testing." Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and not of NSF.

CONTEXT FOR THIS TALK

- Repeated difficulties in teaching software metrics to undergraduate and graduate CS students
- Repeated difficulties in getting bright software engineering academics and professionals to consider issues related to validity, especially construct validity
- Stunning, persistent lack of attention to “the attribute” in software engineering papers on measurement (by practitioners and by academics)

- We suspect that part of the problem is the difficulty of translating too many ideas across disciplinary boundaries.

Preliminary comments: Metrics in Finance

Bank of America

- Assets (book value) per share $19.83
- Price per share $ 7.82
- "Price to Book" Ratio 0.39

- According to these statistics, if you closed B of A and sold its assets, you could get nearly 3x as much as it is worth as a running company.

http://finance.yahoo.com/q/ks?s=BAC+Key+Statistics

CONTRAST

Bank of America	Wells Fargo Bank
• Assets per share $19.83	$25.70
• Price per share $ 7.82	$33.91
• Price to Book Ratio = 0.39	$1.32

What's going on?

http://finance.yahoo.com

Perhaps B of A's "book value" is incredible

- Foreclosed houses – what are they worth?
- How many loans are bad?
- How does this compare to its loan loss reserves?

FINANCIAL RATIOS

- Price to earnings ratio – how much you pay for each dollar of earnings.
- Price to book ratio – how much you pay for each dollar of assets
- Price to sales ratio – how much you pay for each dollar of gross revenue

FINANCIAL RATIOS

P/E, P/S, and P/B are all
widely used by investors,
including well-informed professionals

FINANCIAL RATIOS

Almost no one thinks they are valid

FINANCIAL RATIOS

I don’t think they are valid

FINANCIAL RATIOS

I use them every day

Financial ratios

- P/E, P/S, and P/B are all widely used
- Investors (including professionals) use them every day
- Almost no one thinks they are valid
- Almost no one thinks they are accurate
- I don't think they are valid or accurate
- I use them every day

- What if someone said,

Don't use that!!!

Triangulation: Hunting patterns in weak data

For me, the key to working with a financial ratio is understanding **what that's supposed to tell me about**.

For Price / Book, the underlying concept is how much asset I get for my money. If the company is at risk, this is important.

But if I am actually concerned about that, I look at other indicators of the company's assets and who else has claims against them:

- What potential losses are on the horizon?
- How much do they owe?
- When are those debts payable?
- What challenges have been made to the valuations?
- What history does the company have of surprising revaluations?

Taken together, the collected data might tell a useful story.

"In the land of the blind, the one-eyed man is king."

Desiderius Erasmus (1466-1536)

WHAT ABOUT SOFTWARE METRICS?

"METRICS THAT ARE NOT VALID ARE DANGEROUS."

Cem Kaner

James Bach

Bret Pettichord

Lessons Learned in Software Testing

MANAGERS HAVE LEGITIMATE NEEDS

- They need metrics in order to (for example…)
  - Compare staff
  - Compare project teams
  - Calculate actual costs
  - Compare costs across projects or teams
  - Estimate future costs
  - Assess and compare quality across projects and teams
  - Compare processes
  - Identify patterns across projects and trends over time

- Executives need these, whether we know how to provide them or not.
  - Hung Quoc Nguyen

SOLVE THIS WITH QUALITATIVE MEASURES?

- Maybe, but
  - Expensive
  - Time-consuming
  - Hard to do well
  - Not very good for quantitative comparisons
  - Not very good for modeling
  - Not very good for estimation

- And they suffer from their own quality problems

Note: Expanded diagram from what was presented at CAST 2012

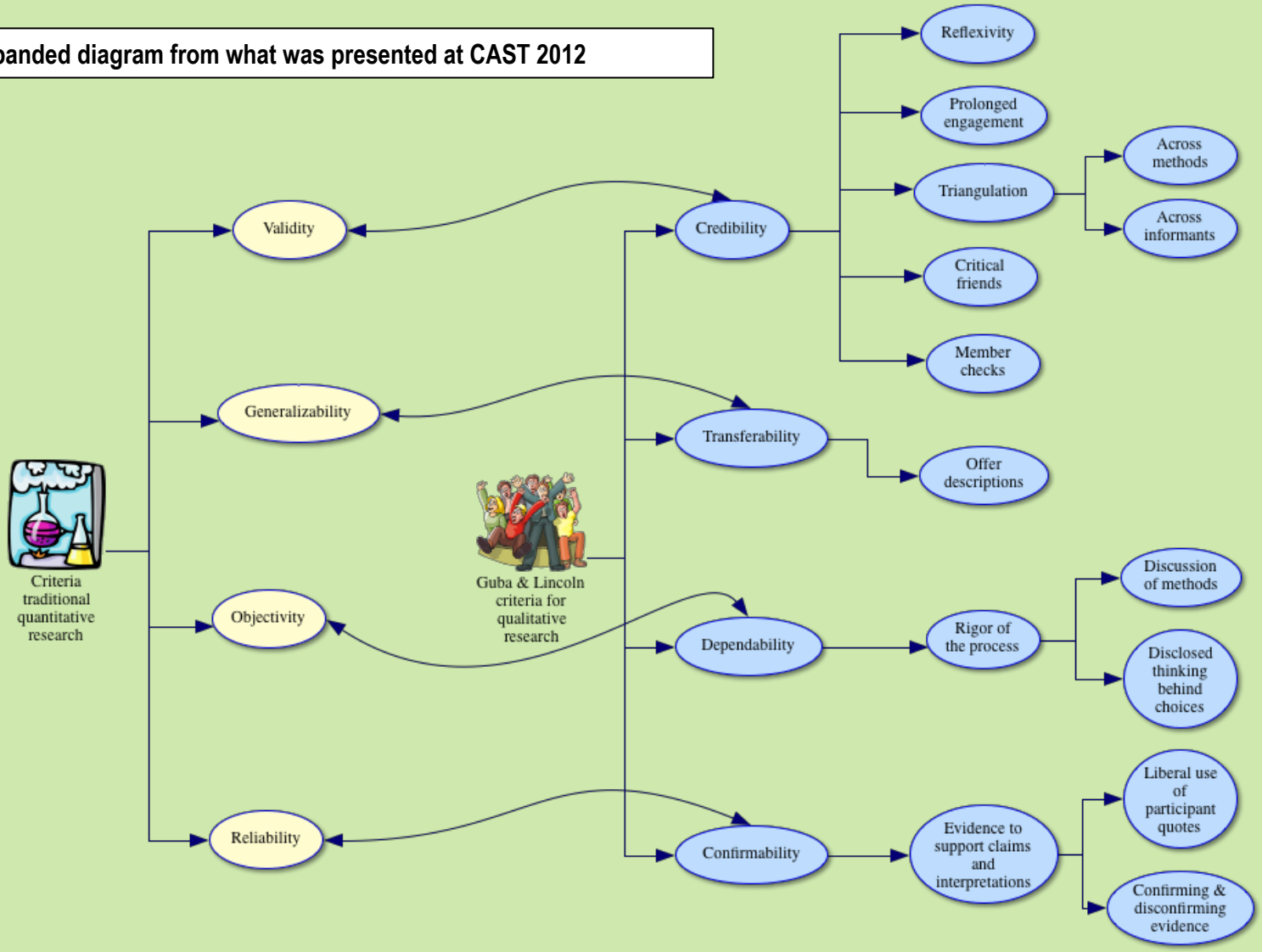


Diagram based on Lincoln, YS. & Guba, EG. (1985). *Naturalistic Inquiry*. Newbury Park, CA: Sage Publications.

QUALITATIVE MEASURES

- Are far from risk-free.
- The compelling story that paints a false picture
  - is no less “bad”
  - than the compelling statistic
  - that gives a false impression

- If metrics that are not valid are **dangerous**
- Then it is important to be aware of their **risks**
- And to manage those risks

- If metrics that are not valid are **dangerous**
- Then it is important to be aware of their **risks**
- And to manage those risks

- As a general rule, testers inform project managers of project risks
- And the managers manage them

It has been 12 years since we published *Lessons*

- And in those 12 years, the field of software metrics has made remarkably little progress.
- Unrecognized and unmanaged threats to validity are still a critical problem
- I am not aware of any collection of validated measures of anything that I want to measure
- We are not on the verge of a solution
- The primary alternative (qualitative measures) carries equal risk
- **This is a hard problem**
- Reasonable people will have to make do with inadequate ways of dealing with this.
- **All** of those ways will carry risk.

LET'S RIDICULE SOME SOFTWARE METRICS

- Defect removal ratio

$$\frac{\text{How many you found}}{\text{Total present in the software}}$$

(Maybe it is more interesting to ask how many of the bugs you missed are surprising.)

- Test case pass ratio

$$\frac{\text{How many tests passed}}{\text{Total number of tests you ran}}$$

(What about all the tests you haven't run? How interesting IS IT that 5000 tests passed out of 5,000,000,000,000,000,000,000,000,000,000,000,000 possible tests?)

These are as bad as Price / Book ratios

So, should we use them?

- I have not personally found them useful
- I do not personally recommend them
- I question their validity on several grounds…

- But managers ask for them
- Defect removal ratio ("defect removal efficiency") is even advocated in several books and papers

Is it unethical to provide them?

- I would not recommend them
- I think you should consider carefully whether you should recommend them (I think you should not.)

- But if someone asks you for them, are you ethically bound to say “no”?
- What is the context?
  - Collect same data across projects, looking for patterns?
  - Collect data as part of a collection of imperfect indicators, looking for patterns?
  - Collect data because an executive or auditor or customer demands them?
- **We can disagree without one of us being “unethical”**

Measurement

The empirical, objective assignment

- of numbers
- to attributes of objects or events
- according to a rule
- derived from a model or theory
- with the intent of describing them.

(Kaner & Bond, 2004)

Measurement

There are no measures in isolation.

A count

- of (bugs, lines of code, dollars, puppies)
- is meaningless as a measure
- until you identify **the attribute**
- that it is intended to measure.

Measurement of an Attribute

The most common problem in software metrics is the

- presentation of a statistic (like bug count)
- without careful and explicit consideration of
- what attribute the statistic is supposed to measure
- and how the value of the statistic is driven by the value of the attribute

In many cases,

- the statistic is presented AS IF IT WERE the attribute (for example, bug count, rather than the underlying attribute, which might be reliability or quality or tester productivity).

This is not measurement

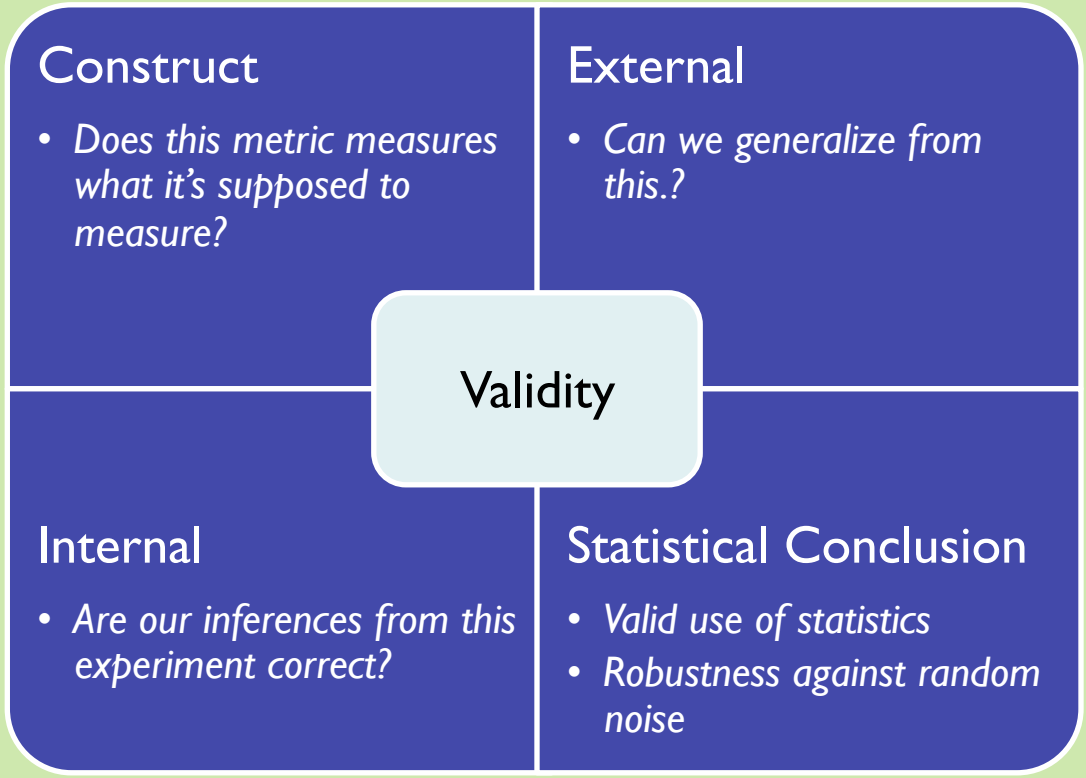
Defining Validity

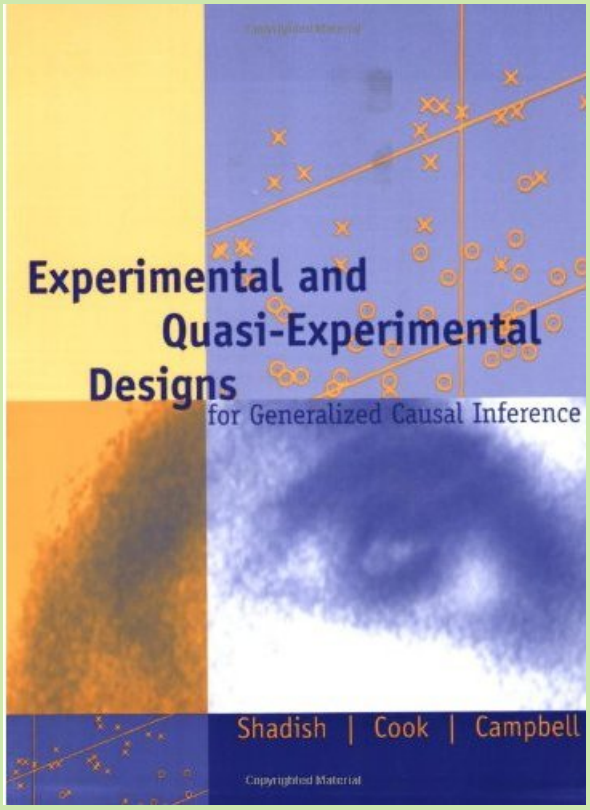
"We use the term validity to refer to the approximate truth of an inference." It's more useful to think of validity as relative rather than absolute. i.e., thinking of degrees of validity, instead of an absolute valid or invalid. (Shadish, Cook, and Campbell, 2001)

Validity is a property of a proposition (Trochim & Donnelly). The literature typically talks about the validity of a conclusion reached through research, but for measurement, the more applicable proposition is that THIS is a good measure of THAT.

A measurement is valid to the extent that it provides a trustworthy description of the attribute being measured. (This is our current working definition).

A Typology of Validity





Shadish, Cook, & Campbell, 2002

Types of validity: From Shadish et al. p. 38

- **Construct validity**: The validity of inferences about the higher order constructs that represent sampling particulars
- **External validity**: The validity of inferences about whether the cause-effect relationship holds over variation in persons settings, treatment variables, and measurement variables
- **Internal validity**: The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured
- **Statistical conclusion validity**: The validity of inferences about the correlation (covariation) between treatment and outcome

APPLY THESE TO MEASUREMENT

- The Shadish, Cook & Campbell taxonomy is focused mainly on the validity of inferences made from the results of experiments.
- Many of those experiments are done to establish the relationship between a statistic and an attribute.
- We aren't talking about the experiments. We're talking about the measurements that are based on the (established or hypothesized) relationships.
- The examples and terminology of Shadish et al are wonderfully suited to the experimental setting, but need translation for measurement.

CONSTRUCT VALIDITY

Constructs: The "ideas" (the abstractions) associated with our measurements.

- Example: Count bugs in order to measure tester productivity
  - The measurement (the statistic) – the bug count
  - The underlying attribute – tester productivity
  - The constructs:
    - **What's a tester?**
    - **What's a productivity?**
    - **What's a bug?**

Construct validity: The extent to which our measurements describe the desired attribute.

Surrogate (or Proxy) Measures

We use a surrogate measure when we

- Don't know what the underlying attribute is
- Don't know how to measure the underlying attribute
- But believe that an easy-to-conduct operation will yield a result that correlates with the value of the attribute
- NO WAY TO GUAGE THE DEGREE OF CONSTRUCT VALIDITY
- HARD TO GUAGE INTERNAL OR EXTERNAL VALIDITY, TOO.

External validity

The extent to which descriptions or conclusions based on the measurements can be applied to other events or objections involving the same attribute.

-- or more simply --

The extent to which we can generalize what we learn from our measurements to other (similar) things or other situations.

CONTRAST CONSTRUCT AND EXTERNAL VALIDITY

Construct Validity

- If a measurement is not tightly focused on the attribute you are trying to study, it will have problems with construct validity.
- If a measurement relies on undefined or unclearly defined concepts (so there is unclarity about what you are counting or what the count's value represents), it lacks construct validity.

External Validity

- If a measurement seems correct, but influenced by the specific conditions under which you took it, it has problems with external validity.

STATISTICAL CONCLUSION VALIDITY

Taking measurements includes statistical (and other mathematical) operations.

These are valid to the extent that they accurately characterize the underlying data.

Internal validity

Taking measurements involves empirical **operations**: We use an instrument and take readings from it.

("Operations": Things we intentionally do)

Internal validity is present to the extent that our operations are sound: we apply the instrument to the appropriate things, apply it competently, and read it accurately.

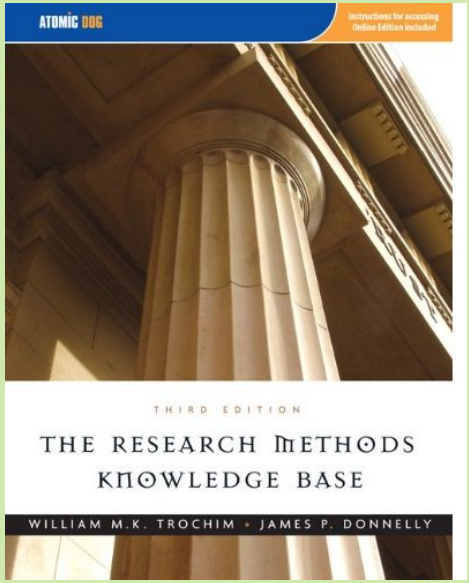
CONTRAST INTERNAL AND STATISTICAL CONCLUSION VALIDITY

- Sources of error that tend to bias the experiment (systematically over-estimate or under-estimate a measurement) are usually problems of internal validity.
- Sources of error that add variability (imprecision; inaccuracy that is not necessarily biased) to the experiment are usually problems of statistical validity.
- Some sources of error add bias AND variability. I often find these hard to classify as statistical versus internal.

Subtypes of Construct Validity

Construct Validity
(Trochim taxonomy)

Translation validity		Criterion-related validity			
Face Validity	Content Validity	Predictive Validity	Concurrent Validity	Convergent Validity	Discriminate Validity



Trochim & Donnelly (2006)

SUBTYPES OF CONSTRUCT VALIDITY

How do you know

that you are measuring

what you think you are measuring?

- **Face Validity**: This measure ***appears*** to be a measurement of that attribute. (Superficially, "on its face", it is a plausible measure).
- **Content validity**: If an attribute is multidimensional (for example, we have considered several factors that are part of an employee's productivity), the measure has content validity to the extent that it represents all of the dimensions.

SUBTYPES OF CONSTRUCT VALIDITY

- **Predictive validity:** Sometimes the value of one attribute should have an impact on the value of another. (For example, more complex programs should take longer to understand.) A measure has predictive validity if its value predicts the value of a measure of the other attribute.
- **Concurrent validity**: If you use two measures of the same attribute of the same things (e.g. measure the length of the same programs in two ways), does this measure correlate with the other measure?

SUBTYPES OF CONSTRUCT VALIDITY

- **Convergent Validity**: Several measures appear to measure this attribute or some aspects of it. To what extent does this measure lead to the same conclusions as those?
- **Discriminant Validity**: If some other construct is different from this one (but sometimes confused with this one), does this measure behave the way it should if it was a measure of our construct and NOT behave the way it should if it were a measure of the OTHER construct?

THREATS TO VALIDITY

- The best discussions of research validity that we've seen (Trochim & Shadish et al) are about validity of research inferences, rather than about measures per se
- Both presentations illustrate the types of validity in terms of **threats to validity.**
- The threats to validity serve as examples of ways in which a given experiment (or measure) may be made less valid.
- In Kaner's experience as a student (and later teacher) of human experimental psychology, working through the threats was often the way that students actually learned the main validity concepts.

CLASSIC EXAMPLES OF THREATS TO CONSTRUCT VALIDITY

Are we measuring what we think we are measuring?

- Inadequately defined constructs (inadequate explication)
- Construct-definitions that omit important dimensions of the construct (mono-operation bias)
- Measurement of the construct in only one way (mono-method bias; the opposite of convergent validity)
- Biases in the conduct of the experiment that cause measurement of the wrong thing or cause other variables to influence the result

These lists of examples are based on Shadish et al. Parenthesized terms are names of the threats used in Shadish's presentation.

CLASSIC XAMPLES OF THREATS TO EXTERNAL VALIDITY

Can we generalize what we have learned from these measurements, to form reasonable expectations of the measurements we would obtain under other circumstances?

- Effects might be specific to settings (what happens at Intel might not happen at Electronic Arts)
- Effects might be specific to people (is there anything "special" about this group of people, this manager, or this experimenter)?
- Effects might be specific to this type of project (rocket science might be different from web databases)
- Effects might be general but with exceptions (works with everything BUT rocket science)
- Overfitting of data or overoptimizing of the experiment achieves results that cannot be generalized

CLASSIC EXAMPLES OF THREATS TO STATISTICAL VALIDITY

Are our data analyses sound and do they justify the conclusions we are drawing about the meaning of the data?

- Violated assumptions of the statistical model
- Computational inaccuracy (rounding error, or poor algorithm)
- Unreliability of measures
- Restriction of range
- Excessive impact of extreme values
- Extraneous variation in the setting in which the measurement is taken

We often look for correlations between the variable we are studying and other variables that we believe are also tied to the underlying attribute. All of the statistical problems described by Shadish et al. apply to this work.

CLASSIC EXAMPLES OF THREATS TO INTERNAL VALIDITY

Are our measurement-related operations sound?

In particular, are they free of bias?

- **Selection effects**. For example, picking hard-to-maintain programs that also have characteristics that will generate high complexity estimates. Are there easy-to-maintain programs with the same characteristics?
- **Maturation and History**: Other aspects of the project (including the skills of the staff) change over time and so what looks like a change in one variable might actually be the result of a change in something else
- **Instrumentation and human operations**: For example, what tools or procedures you use to count lines of code. If the tool merely increases measurement error, it creates a statistical validity problem. If it creates bias, it is an internal validity problem

A CHALLENGE OF TERMINOLOGY

- Well before this point, my students are typically hopelessly lost in the terminology
- This terminology is not widely used in the computer science literature
- In fact, even “construct validity” almost never appears in discussion of experiments or metrics.

A VARIATION OF THE TAXONOMY

- To try to improve our students' comprehension (and perhaps to increase usage of the concepts across the field), we are trying out an adaptation
  - Renaming parts of the taxonomy
  - Redescribing several of the threats, with computing-related examples.
- We are loathe to invent an alternative taxonomy
- This is an initial working model
- If it doesn't work for our students, we'll drop it

OUR PROCESS

We (mainly Nawwar) brainstormed a collection of 64 threats to validity

- Based heavily on Shadish et al (2002), Trochim & Donnelly (2006) and Austin (2006)
- Supplemented by our experiences

VALIDITY TYPES

- **Construct**: How well our concepts are defined and how we use them to design, control and interpret our measurements
- **Generalization** (external validity): How well we can generalize from these measurements to expected measurements of other things, at other times
- **Operational**: How well we take the measurements and handle the data
- **Purpose**: We take measurements for a purpose. Will this measure actually help us achieve this goal?
- **Side-Effect:** Taking a measurement changes the system being measured. It impacts the relationship between the operations and the attribute. How well this is managed.

Examples of Threats to Construct Validity

How well our concepts are defined and how we use them to design, control and interpret our measurements

EXAMPLES: THREATS TO CONSTRUCT VALIDITY #1

- **No construct**: Measurement is attempted without an explicit attribute. We have no idea what we are actually trying to measure.
- **No model**: There is no clear or well-justified theory or model that justifies the assumption that the metric and the attribute are related, or related in the assumed way.
- **Poorly understood attribute**: The measurement is tied to a vaguely-understood or vaguely-defined construct.
  - Example, what is “productivity”?

EXAMPLES: THREATS TO CONSTRUCT VALIDITY #2

- **Likely to be misunderstood**: The measurement is well-tied to its intended construct but will be misinterpreted by our audience as tied to a different construct.
  - Example: A task is designed to assess the maintainability of the code (looking at the likelihood of adding new bugs when someone changes the code) but the people who interpret the data will interpret the results as measurements of the program's reliability.
- **Misdirected**: The attribute is misconceived.
  - Example, we perceive a task as reflecting on the skill of team when it is more closely tied to the process they are following.

EXAMPLES: THREATS TO CONSTRUCT VALIDITY #3

- **Ambiguous**: The construct (the attribute or a measurement construct) is ambiguously defined.
  - Example: if the measure counts the number of "transactions per second", what is a "transaction"? (Will different people count different things as transactions? Or the same person at different times? How much variation in borderline cases?)
  - Example: Count bugs. Include "enhancements"?
- **Confounded**: The measurement is jointly determined by the value of the attribute and of some other variable(s).
  - Example, a tester's rate of reporting bugs is partially related to the tester's skill but also related to the program's bugginess.

EXAMPLES: THREATS TO CONSTRUCT VALIDITY #4

- **Definition too narrow** (mono-operation bias): The underlying attribute is multidimensional but our measurement is tied to only one or a few of the attribute's dimensions.
  - Example: consider measuring a tester's skill only in terms of her bug reports, ignoring her skills in coaching, design, and automation development).

EXAMPLES: THREATS TO CONSTRUCT VALIDITY #5

- **Measurement too narrow** (mono-method bias): We measure the construct in only one way (this is the opposite of convergent validity)
  - Example: Imagine assessing the maintainability of a programmer's code.
    - We could do a code review for maintainability
    - We could try to change the code and see how difficult it is and how many surprising bugs get introduced
    - We could give the code to someone who doesn't know it and time how many hours they need before being willing to say they understand it.
    - There are many ways to assess the maintainability, none perfect. If we pick only one, our assessment is likely to be less accurate (and might be biased) compared to combining measures.

EXAMPLES: THREATS TO CONSTRUCT VALIDITY #6

- **Surrogate measure**: We decide that using a measurement that is closely tied to the underlying attribute is too hard and tie the measure to something that is (probably) correlated with the underlying attribute instead.
- **Bias:** Biases in the planning or conduct of the experiment that cause measurement of the wrong thing or cause other variables to influence the result. This is similar to a surrogate measure, but not the same.
  - In the case of the surrogate, you intentionally choose an indirectly-related measure, knowing that it is only a correlate of what you want to study.
  - In the biased case, you think you're doing the right thing. Your biases affect your judgment or your perception of the variable or of your activities.

Examples of Threats to Generalization (External Validity)

How well we can generalize from these measurements to expected measurements of other things, at other times

Examples: Threats to Generalization #1

- Effects might be specific to organizations
  - What happens at Intel might not happen at Electronic Arts.
- Effects might be specific to settings or environments
  - What happens in India might not happen in Canada.
  - What happens in the development lab of medical-device maker with expensive instrumentation and subject-matter-expert staff might not happen in the lab of a consumer-product company.
- Effects might be specific to people
  - There might be something "special" about this group of people, this manager, or this experimenter

Examples: Threats to Generalization #2

- Effects might be specific to this type of project
  - Rocket science is different from web databases
- Effects might be specific to the size of the project
  - Projects with 10 people have different dynamics than projects with 1000
- Effects might be general but with exceptions
  - A development process might work in most environments but fail for large national-defense (classified, no-foreign-staff) projects.

Overfitting of a set of data or overoptimizing the design of an experiment achieves results that are often specific to the data set or the experimental context. The results often cannot be successfully generalized.

Examples of Threats to Operational Validity

How well we take the measurements and handle the data

Examples: Threats to Operational Validity (#1)

There are so many of these that we think we need to subdivide them. Here are our current subcategories:

- **Sampling error**
  - Collect data that doesn't accurately reflect the population
- **Random error in observation or recording**
  - This is classical measurement error. The measured value differs from a hypothesized underlying true value by a randomly varying amount
- **Biased observation or recording**
  - Our measured value is systematically too low or too high
- **Analysis error**
  - The collected data are handled improperly or subjected to mathematically inappropriate analyses

Examples of Sampling Error (Operational #2)

- Measure tasks that a programmer has experience with. Generalize to his overall productivity.
- Count bugs in part of the program. Extrapolate to the entire program.
- Measure the performance of an individual rather than of the overall group the individual works in.
- Measure (combine data from) a group rather than segregating data by individual.
- Measure an individual against peers when a better measure would be against the individual's own past work (e.g. improvement scores)

Examples of Sampling Error (Operational #3)

- Count together items that are essentially different or count as distinct groups items that are essentially the same.
- Measure an individual or group against incomparable others. (e.g. compare university students to experienced professionals.)
- Measure too few distinct samples (e.g. evaluate work done only in January, when the individual was sick, rather than sampling across projects, months, etc.)
- Measure only people or items that are still around after certain events or task/life stages. (This is often called "attrition." We assume that the people who left are equivalent to the people who stayed. Example: measure only people who are still employed after rounds of layoffs, but assume their work will be equivalent to an "average" group that hasn't gotten rid of its weakest performers.)

Examples of Sampling Error (Operational #4)

- The people being measured change what they do as a result of other events during the period of the experiment, e.g. gradually gaining experience over a 2-year study or learning from some unrelated historical event. The bias comes in attributing this to a change caused by the experimental conditions.

Examples of Measurement Error (Operational #5)

- Basic measurement error: Our instrument is imprecise or what we do when we use the instrument is imprecise. (Try measuring the length of a multiple-mile race track in inches or the length of a months-long project in seconds.)
- Count together items that have significantly unequal variance or differently-shaped distributions (that is, count them as if they were of the same type)
- Truncating outliers without an appropriate and justifiable theory of extremes
- Not truncating outliers that are actually intrusions from a different process with an fundamentally different distribution

Examples of Measurement Error (Operational #6)

- Miss items because of lack of visibility (they are hard to see or easy to not notice)
- Count duplicates as if they were different (This is measurement error if due to random sloppiness rather than to bias)
- Count duplicates as if they were different because they appear to be different (Example: multiple reports of the same bug, each describing slightly different symptoms)
- Memory error. Record the data well after an event has happened, relying on someone's memory. The delay introduces a random error (misremembering).
- Criterion (or construct) variation. Assign values to observations based on criteria that vary within and/or between observers.

Examples: Errors caused by bias (Operational #7)

Think of a bias as an influencer that introduces error into your measurements (or your interpretation of your measurements) in a consistent direction (e.g. too high or too low). A source of bias can be inanimate. It doesn't necessarily reflect a bad attitude or a bad person.

- Lack of visibility: Miss items because you (systematically) don't see them.
  - Example: your bug counts include things reported to Tech Support, but only the ones they tell you about
- Reporting disincentives: Miss items because there are pressures in the system you are studying against reporting them.
  - Example: Managers force staff to work overtime but report only 40 hours

Examples: Errors caused by bias (Operational #8)

- Reporting disincentives: Underestimates or overestimate because of pressures in the observed system. For example:
  - A group systematically files multiple reports of the same thing (e.g. policy to file one bug report for each variation in observable symptoms of the same bug)
  - A group systematically groups reports as duplicates when they are not (e.g. over-aggressive policy to purge duplicates from a bug reporting database)
  - Ignored or misreported time because of a bias against spending or reporting time on that activity (e.g. time spent on administrative activities, coaching, or attending meetings or training)
  - A resource is used for several purposes (maybe simultaneously) but the entire use is tracked against one purpose, or usage is double-counted
  - Counting resource use when it was not used or not counting resource use when it was used (e.g. holiday / weekend)

Examples: Errors caused by bias (Operational #9)

- Data collected for a different purpose systematically misses information relevant to this purpose
  - Example: Recent trend to study development process via post-hoc content analysis of the code changes and associated check-in comments in a source code control system. What information was NOT entered in those comments?
- Resource use estimated well after the fact, introducing errors and misallocations due to memory effects
- Measure is dominated by evaluation-related skill rather than by the underlying attribute. (Related phenomenon in psychology is discussion of performance versus competence). Example: Two different versions of a status report for the same events might create different impressions of how resources were used.

Examples: Errors caused by bias (Operational #10)

Unintended effects of the measurement activities or experiment itself

- Hawthorne Effect: Introducing a change to a system (including a new measurement) can create enthusiasm. They work harder or more carefully for a short period, and your measured results look better. You attribute it to the systematic change, not the transient enthusiasm.
- Novelty Effect: Introducing a change may temporarily disrupt underlying performance until people get used to it.
- Compensatory Rivalry: Introducing a new method or tool to one group might create competition (such as unreported voluntary overtime) from other groups who want to show they are “just as good”.

Examples: Errors caused by bias (Operational #11)

Demand Characteristics: Aspects of the situation influence people who generate the data you are recording in ways that influence them to change what they do or report in order to give you what they think you want. Examples:

- Organization tracks when bugs are reported and fits them against a theoretical bug-rate curve. To please management, people delay reporting some bugs in order to keep the numbers close to the curve before a project milestone.
- Data collected for a different purpose systematically misses results relevant to this purpose (e.g. what gets missed when we do a post-hoc content analysis of check-in comments in a source code control system)
- People take extra care to ensure that some types of results are reported, resulting in duplicates.
- Systematic filing of multiple reports of the same thing (e.g. policy to file multiple reports of the same bug, one for each variation in observable symptoms)

Examples: Errors caused by bias (Operational #12)

Experimenter effects: The experimenter (or measurer) (you) influences the situation in ways that will influence the measured results.

- Carefully check results that you regard as implausible or undesirable but take the other results at face value.
- Refine your operations when they generate unexpected or undesired results but leave alone equally-rough operations that yield expected or desired results
- Discard legitimate data that you perceive as outliers
- Don't discard (don't look for, don't notice) genuine outliers that drive averages in the direction you want or expect
- Tell people (or otherwise convey) your measurement goals in ways that set up demand characteristics

Examples: Errors caused by bias (Operational #13)

Experimenter effects (2):

- Miss items because of observer bias
  - Don't notice them
  - Specifically don't notice (don't look for) low-probability items or disbelieve that they occurred
- Use weights for values or variables without a sound theory underlying the weighting scheme

Examples: Mathematical or statistical errors (Operational #14)

- Incorrect scale
  - Example: Treat ordinal numbers as interval (incorrectly assume that the amount of distance between the numbers is meaningful)
- Incorrect arithmetic operation for a number of this scale
  - Example: Compute averages of ordinal numbers or of numeric labels
- Lack of statistical power
  - A correlation or experimental effect might appear statistically small because the sample size is too small or the statistical test is too weak. In either case a conclusion of no effect or no correlation is mistaken.
- Using invalid statistical operations (such as confidence bands based on an incorrect assumption of distributional symmetry)
- Adopt a model based on curve-fitting, even though the underlying assumptions don't work, and then force future data to the model (interpret future data in terms of the model).

Examples of Threats Associated with the Purpose of the Measurements

We take measurements for a purpose. Will this measure actually help us achieve this goal?

Examples: Threats related to purpose #1

The overall risk is that the measurement process doesn't help us achieve the goal of measurement.

This section is incomplete and perhaps should merge with the section on distortion and dysfunction (next). But we are keeping them separate in order to preserve mis-measurement here, distinct from mischief caused by (and changing) the measurement (next).

EXAMPLES: THREATS RELATED TO PURPOSE #2

Measure an attribute that is only a small part of what is relevant to the goal. For example:

- To decide who to lay off, assess only "productivity" and ignore other attributes of good staff such as "quality of work" and "teamwork" and "ability to create original solutions"

Measure an attribute that may or may not be well related to the goal. For example:

- Bug counts used to measure quality carry different risks for these objectives (a) assess progress relative to schedule, (b) assess productivity of testers, (c) assess productivity or skill of programmers, (d) compare development processes in terms of productivity, (e) compare development organizations.

Examples of Threats Associated with Side Effects of the Measurement

Taking a measurement changes the system being measured. It impacts the relationship between the operations and the attribute. How well this is managed.

Key reference: Austin (1996)

Examples: Threats related to side-effects #1

The underlying theme is that people change what they do in response to how they are measured. This is normal. We expect this. It is the basis of measurement-based management.

But if people have finite time, and they give you more of what you measure, where do they cut back? What do you lose in order to obtain these perceived gains?

- **Measurement distortion**: An effect of taking these measurement is to change the system in ways that are undesirable. Example: reallocate resources in ways that starve an unmeasured task
- **Measurement dysfunction**: The measurement distortion is so bad that the system-under-measurement looks better than before measurement but is actually worse than it would have been without measurement.

Examples: Threats related to side-effects #2

- People might color or falsify the data you get
- People might stop doing important but unmeasured tasks
  - Managers reassign people who persist in important but unmeasured tasks or who persist in reporting undesired results
- Counts of undesirable things go down because activity is reduced, not improved. Examples:
  - Do less testing because you are writing more status reports: find and report fewer bugs.
  - Delay a critical kind of testing until after a milestone or until someone goes on vacation
- People delay or underreport "undesirable" things (causing consequences of not dealing with them)

Examples: Threats related to side-effects #3

- People might create problems that they can then get credit for fixing
- Emphasis on individual performance and productivity can reduce
  - Coaching
  - Collaboration
  - Time invested in building tools, especially tools used by the broader group
- People might increase measured activities in ways that introduce unmeasured risks
  - Examples: drive to greater apparent productivity can wear out equipment from overuse, burn out people, yield errors, cause accidents

REFERENCES

- Robert Austin (1996), *Measurement and Management of Performance in Organizations*.
- Yvonne S. Lincoln & Egon G. Guba (1985) *Naturalistic Inquiry*
- Doug Hoffman (2000), "The Darker Side of Software Metrics", http://www.softwarequalitymethods.com/Papers/DarkMets%20Paper.pdf.
- Cem Kaner & Walter P. Bond (2004), "Software engineering metrics: What do they measure and how do we know?" http://www.kaner.com/pdfs/metrics2004.pdf
- Michael Q. Patton (2002, 3rd Ed.). *Qualitative Research & Evaluation Methods.*
- William R. Shadish, Thomas D. Cook & Donald T. Campbell (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.*
- William Trochim & James P. Donnelly (2006, 3rd Ed.) *The Research Methods Knowledge Base*