Deception Dangers of the Numbers Game

LYNN MCKEE

Calgary Software Quality Discussion Group

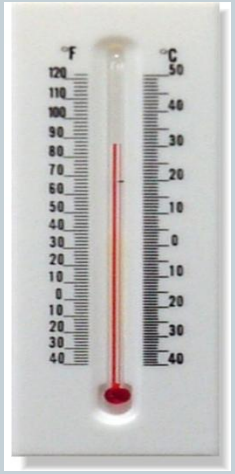
Calgary, Alberta

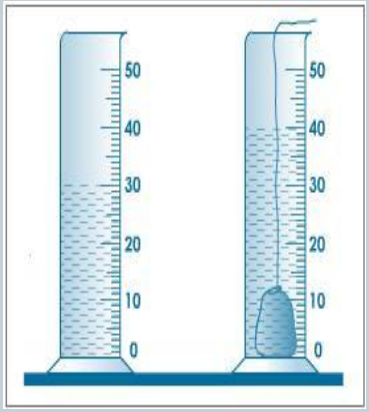
 @lynn_mckee #testing #metrics #sqdg

Agenda

- The Measurement Problem
- Measurement & Context
- Positivism & Post Positivism
- Qualitative vs Quantitative Debate
- Abstraction & Construct Validity
- Orders of Measurement
- Measurement Side Effects
- Contextual Conversation
- Summary

What is Measurement?





Measurement

- What is Measurement?
  - "Measurement is the assignment of numbers to objects or events according to a rule derived from a model or theory."

    - Cem Kaner

  - "The art and science of making reliable observation."

    – Gerald M. Weinberg

  - Michael Bolton uses Jerry's definition and adds "Implicit is the notion of comparison for the purpose of making a distinction."

Common Testing Measurements

Group Exercise

The Measurement Problem

- Distortion and dysfunction are pervasive
- Many things that we would like to measure are subjective
  - complex, qualitative, non-repeatable, and involve human judgment or human performance
- Models are not transferrable
- Context is often more important than the metrics

"Decisions about quality are political and emotional, based on discussions and decisions about whose values count and how much they count relative to one another."

– Gerald M. Weinberg, Quality Software Management, Volume 1: Systems Thinking

Why do we Measure?

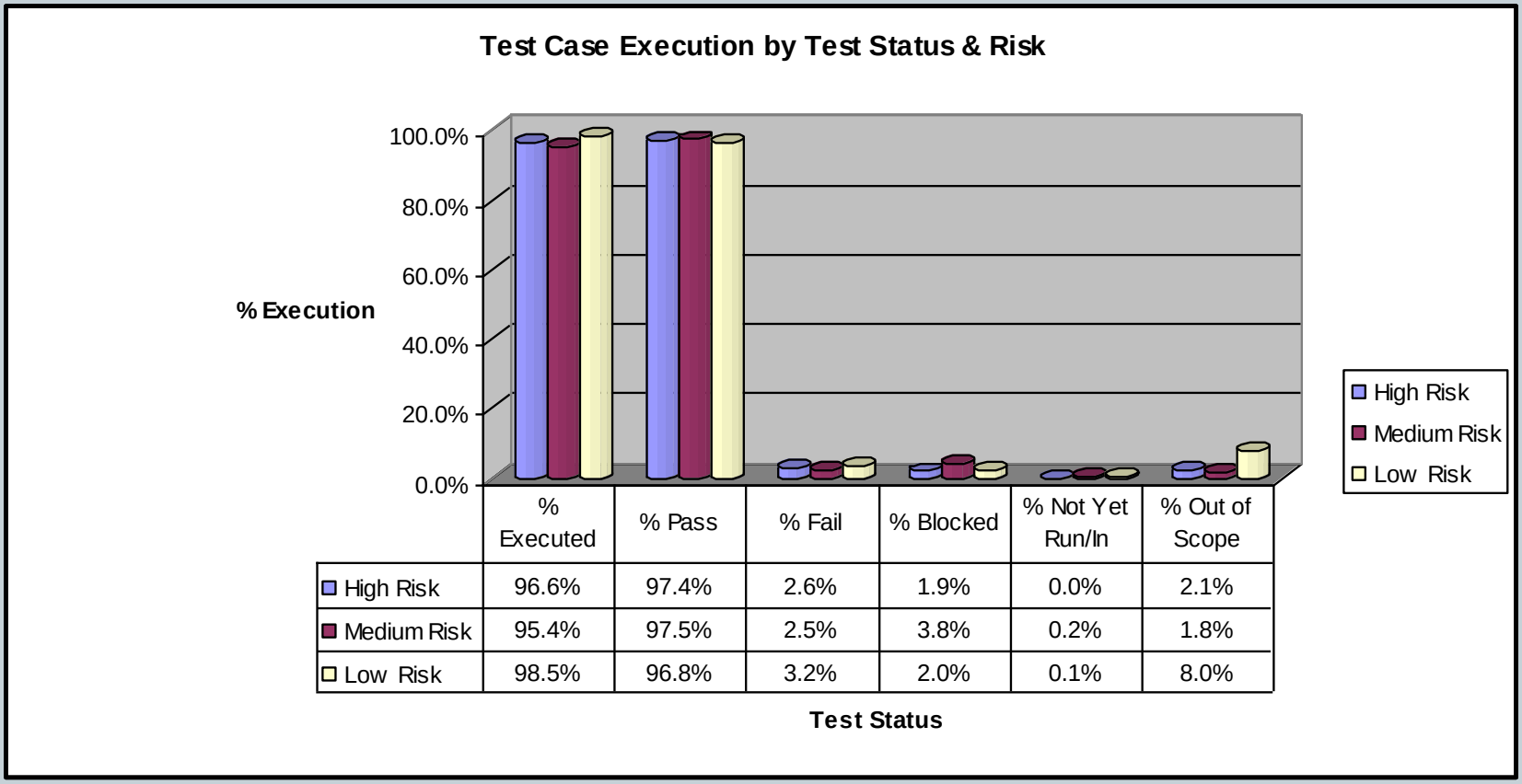
- Some of the many reasons may include:
  - Track project progress
  - Gain control of processes
  - Demonstrate the productivity of your staff
  - Demonstrate the quality of your work
  - Compare different practices
  - Increase your credibility with your management
  - Identify where improvements are needed
  - Determine (relative) complexity or other attributes of the software
  - Assess quality levels (value on some desirable attribute, such as reliability, performance, usability, accessibility, etc.)
  - Gain control of characteristics of the products you make
  - Gain the respect of your customers
  - Demonstrate the effectiveness of the product

- Cem Kaner, Measurement Issues & Software Testing

Simplifying Why We Measure

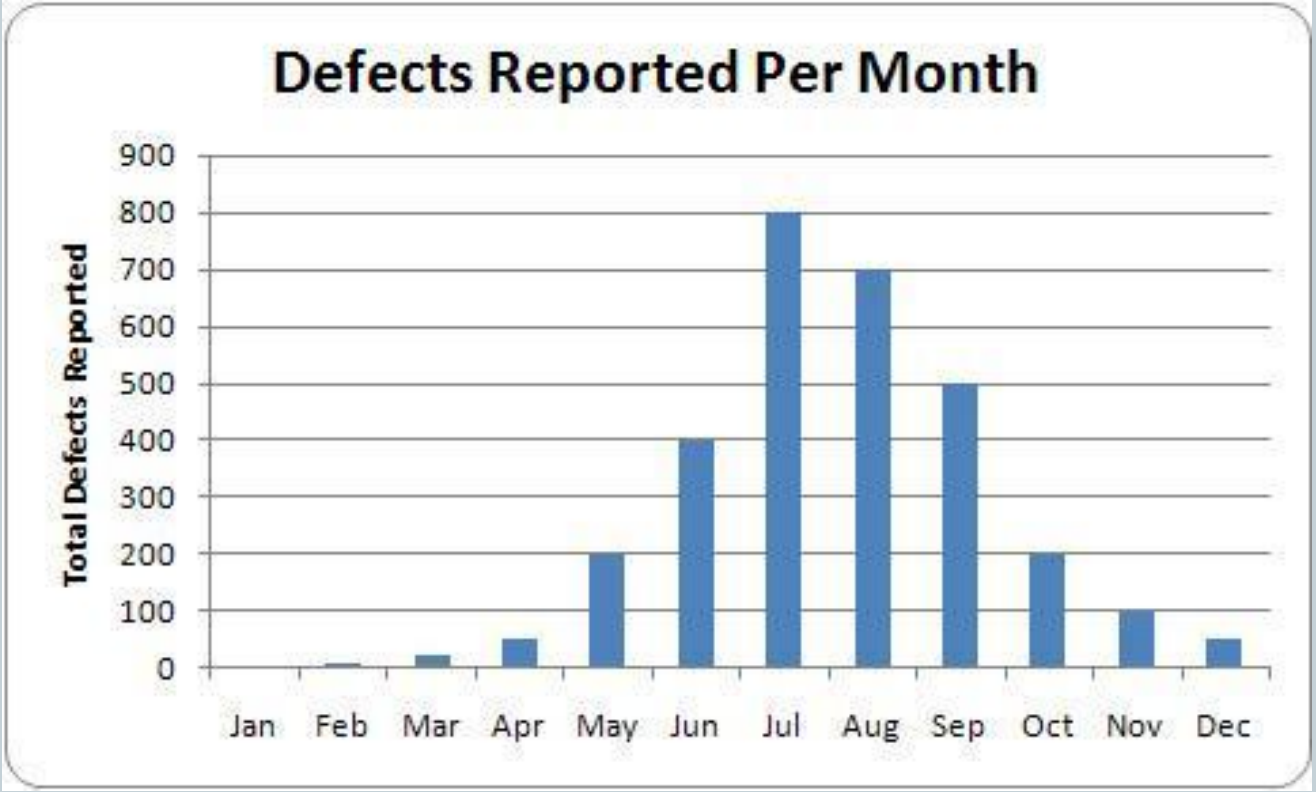
- The reason we measure is to find out if we got is what we wanted
- This cannot be done effectively with numbers alone
- Must compare what you got with what you wanted
- With software this is often a question of “Compared to what?”

Valuable or Not Valuable?



	% Executed	% Pass	% Fail	% Blocked	% Not Yet Run/In	% Out of Scope
High Risk	96.6%	97.4%	2.6%	1.9%	0.0%	2.1%
Medium Risk	95.4%	97.5%	2.5%	3.8%	0.2%	1.8%
Low Risk	98.5%	96.8%	3.2%	2.0%	0.1%	8.0%

Valuable or Not Valuable?



Measurement & Context

- The important of context dates back to Aristotle
- Aristotle argued against using the arithmetic mean in the example: How much should an athlete eat?
- Aristotle responded…

  "It would be absurd to infer from the fact that 10 lbs. is too much and 2 lbs. too little for me that I should eat 6 lbs."

"Finding the mean in any given situation is not a mechanical or thoughtless procedure, but requires a full and detailed acquaintance with the circumstances."

– Aristotle

Is Testing Stuck in Positivism?

- Positivism is a rejection of metaphysics (the nature of being and the world); viewing the purpose of science is simply to stick to what we can observe and measure
- Identifies science as the way to get at truth, to understand the world well enough so that we might predict and control it
- Believes in *empiricism* which emphasizes knowledge that is founded in evidence rather than reasoning, intuition or revelation

- Social Research Methods, http://www.socialresearchmethods.net/

Shifting Testing to Post Positivism

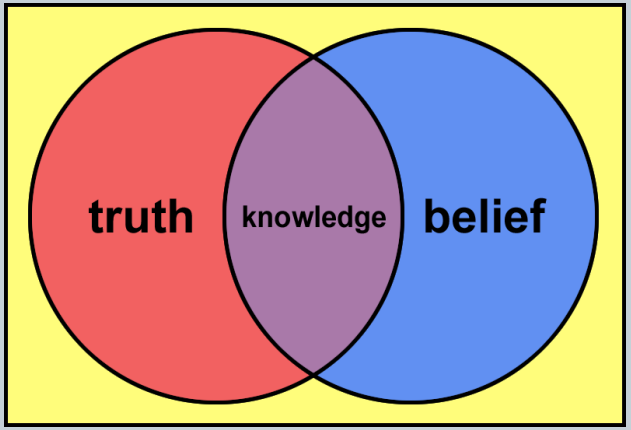
- Post Positivism recognizes that all observation is fallible and has error and that all theory is revisable (Critical Realists)
- Emphasizes the importance of multiple measures and observations; each may possess different types of error
- Identifies the need to use triangulation across these multiple errorful sources to get a better handle on reality
- Recognizes all observations are based on models
- Highlights we are inherently biased by our cultural experiences, world views, etc.

Qualitative-Quantitative Debate

- Qualitative data typically consists of words while quantitative data consists of numbers
- Both have been used to address diverse research topics
- Combining both methods is referred to as a "mixed methods" approach

Qualitative-Quantitative Debate Cont.

- Numbers alone cannot be interpreted without understanding the underlying assumptions
- These assumptions are epistemological and ontological
  - Epistemology is concerned with the nature and scope of knowledge
  - Ontology is concerned with nature of reality and existence



The Qualitative-Quantitative Debate Cont.

- Common myths about the differences:
  - Quantitative research is confirmatory and deductive in nature
  - Qualitative research is exploratory and inductive in nature
- Quantitative can be exploratory and inductive and qualitative can be confirmatory and deductive
- Quantitative excels at summarizing large amounts of data and reaching generalizations based on statistical projections
- Qualitative excels at "telling the story" from the participant's viewpoint, providing the rich descriptive detail that sets quantitative results into their human context

Abstraction

- Models for deriving measures employ the concept of abstraction
- Abstraction retains only the information relevant for a specific purpose
- We need to question what we are *not* seeing when looking at the results of a given measurement model

Construct Validity

- Validity is the best available approximation to the truth of a given proposition, inference, or conclusion
- When assessing validity, the first question needs to be “Validity of *what*?"
- Measures, samples and designs don't have validity
- Only a proposition, inference or conclusion can have validity

Threats to Construct Validity

- Do your measures reflect what you wanted them to reflect?
- How do you know?
- How will you be criticized if you make these types of claims?
- How might you strengthen your claims?
- The kinds of questions and issues your critics will raise are what is meant by threats to construct validity

Threats to Construct Validity Cont.

- Inadequate Preoperational Explication of Constructs
  - Ineffectively defining your construct; requires more time thinking through your concepts and articulating them better
- Mono-Operation Bias
  - Construct is limited to a single variable
- Mono-Method Bias
  - Construct is limited to a single measure
- Interaction of Different Treatments
  - Construct is affected by other programs
- Interaction of Testing and Treatment
  - This concerns the affect both the tests and the measurements have on the attribute being measured

- Cook and Campbell, D.T. Quasi-Experimentation: Design and Analysis Issues for Field Settings

Threats to Construct Validity Cont.

- Restricted Generalizability Across Constructs
  - This threat involves the "unintended consequences" such as serious side effects of the measurement program
- Confounding Constructs and Levels of Constructs
  - Slight increases or decreases to the attributes being measured may dramatically affect the results and being unfair representations
- Hypothesis Guessing
  - The participants of the measurement program adjust their behaviour based on what they guess the objectives of the program to be
- Evaluation Apprehension
  - People are anxious about being evaluated and this can make them perform poorly or even better in attempt to "look good" or "look smart"
- Researcher Expectancies
  - The conscious and unconscious bias by the person acquiring the metrics

Orders of Measurement

- Consider three broad categories of measurement
  - 1st Order
    - Tend to be qualitative, fast and inexpensive
  - 2nd Order
    - More quantitative, subject to more refined models and more involved
  - 3rd Order
    - Precise, high quantitative measures that tend to be about simple systems or simple models

- Gerald M. Weinberg, Quality Software Management, Vol. 2: First-Order Measurement

Orders of Measurement Cont.

- Testing measurements should revolve around 1st and 2nd order
- 3rd order measurement is impossible for software testing
- Focus on what question we are trying answer
- Generate simple measures that generate interesting discussion

Measurement Side Effects

"Before we can assign numbers to our observations, we must understand the process by which we obtained them in the first place."

– Gerald M. Weinberg

A Measurement Story…

Merlin the Manager was tired of being chastised by his boss, Wanda, for low programmer productivity. "How can I show you that the programmers are doing something," he asked her, "when all they're ultimately producing is ones and zeros."

"I'm not interested in zeros," Wanda complained. "Zeros are nothing. How many ones are they producing?“

"Um, I don't know," Merlin stammered.

"Well, you're their manager," Wanda accused. "You should know.“

"Of course," Merlin apologized, backing out of Wanda's office. "I'll institute a metrics program."

Continued….

- Gerald M. Weinberg, Parable of Ones Blogpost

A Measurement Story Cont….

Merlin then hired some measurement consultants who showed him how to count the ones automatically in every object program, plotting them by project and programmer. The initial report showed an overall productivity of 43.78% ones, and Merlin called a meeting of all the programmers to chastise them about their low productivity.

"Look at this figure," he accused. "This means that 56.22% of all bits on memory are essentially unused—filled with zeros. Why, when I was a programmer, I could generate programs at random that were 50% ones. If this keeps up, there won't be any performance awards this year, I can assure you.

Two months later, just before the performance awards were decided, Merlin looked at his metric report and was delighted to discover that the overall productivity figure was 53.04% ones. He showed this report to Wanda, who gave him a big bonus. "Well," he thought, "that certainly shows the value of a measurement program. Now, as soon as I fire those two programmers with less than 45% ones, productivity will show another boost."

Predictable Behaviour of Measurement

- People tailor their behaviour to things that they are measured against
- Behaviours change in predictable ways to provide the answers management asks for
- The desire to measure the productivity and quality of our work is pervasive…
  and dangerous

Measures Requested are Measures Attained

- Consider the example of measuring individuals by their bug reports
- If you ask a tester for more bugs, you'll probably get more bugs.
  - You are likely to get more bugs that are minor, or similar to already reported bugs, or design quibbles -- more chaff. (Weinberg & Schulman, 1974)
- People know that other people tailor their behaviour.
  - Put a tester under incentive to report more bugs and every programmer will become more skeptical of the value of the bug reports they receive.
  - Bug counting creates political problems (especially if you also count bugs per programmer).
- The measurement system creates incentives for superficial testing and against deep tests for serious underlying errors.
  - Bug counts punish testers who take the time to look for the harder-to-find but more important bugs.

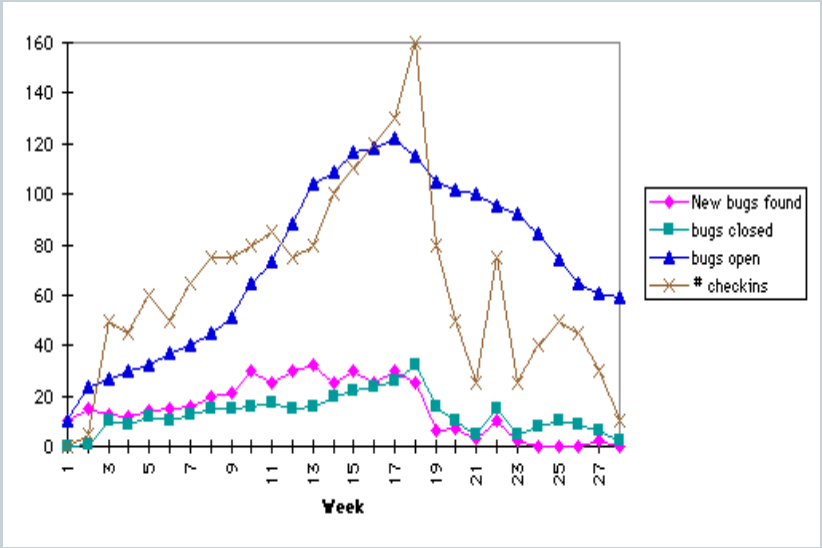
- Doug Hoffman, The Darker Side of Metrics

Measures Requested Continued…

- You can make a tester look good or bad just by choosing what type of testing she should do / what area to test.
  - Regression testing often yields fewer bugs than exploratory testing of the same area of the program.
  - Fewer bugs to find in less buggy code. If raises and promotions are influenced by bug counts, project assignments will often be seen as unfair. More political problems.
- The system also penalizes testers who support other testers.
  - It takes time to coach another tester, to audit his work, or to help him build a tool that will make him more effective. The tester who does this has less time to find bugs.
- Time spent on any process that doesn't lead to more bugs faster is time that counts against the tester.
  - For example, bug counting rewards testers who minimize the time they spend documenting their test cases.

Contextual Conversation

- Metrics need to be used to drive inquiry rather than to control
- Inquiry explores context
- Use numbers to illustrate stories providing rich descriptive detail that sets quantitative results into their human context
- Be wary of numbers becoming placeholders for stories



Contextual Conversation Cont.

- Question measurements by asking:
  - Who Says So?
  - How do they know?
  - What's missing?
  - Did somebody change the subject?
  - Does it make sense?

- Darrell Huff,
How to Lie with Statistics

Summary

- Be mindful that measurement is difficult as what we seek to measure is subjective -- complex, qualitative, non-repeatable, and involves human judgment or human performance
- Combine diverse qualitative and quantitative measures and triangulate
- Question measurement models including the abstractions and construct validity
- Need to be critical of our ability to know reality with certainty; consider the epistemological and ontological assumptions
- Seek simplicity using 1st and 2nd order measurements
- Focus on measurements that generate meaningful discussion
- Think carefully about the potential side effects of your measures

Questions?

Lynn McKee
Quality Perspectives
Calgary, AB Canada
lynnmckee@qualityperspectives.ca
www.qualityperspectives.ca
lmmckee @lynn_mckee lynnmckee

References

- Quality Software Management, Vol. 1: Systems Thinking, Gerald M. Weinberg
- Quality Software Management, Vol. 2: First-Order Measurement, Gerald M. Weinberg
- How to Lie with Statistics, Darrell Huff
- Software Engineering Metrics: What Do They Measure and How Do We Know? http://www.kaner.com/pdfs/metrics2004.pdf
- Negligence and Testing Coverage by Cem Kaner, http://www.kaner.com/pdfs/negligence_and_testing_coverage.pdf
- The Impossibility of Complete Testing by Cem Kaner, http://www.kaner.com/pdfs/imposs.pdf
- Three Kinds of Measurement and Two Ways to Use them by Michael Bolton, http://www.stickyminds.com/s.asp?F=S15136_ART_2

References

- Measurement Issues & Software Testing by Cem Kaner http://www.kaner.com/pdfs/measurement_segue.pdf
- The Darker Side of Metrics by Doug Hoffman, http://www.softwarequalitymethods.com/Papers/DarkMets%20Paper.pdf
- Parable of Ones by Gerald M. Weinberg http://secretsofconsulting.blogspot.com/2010/12/parable-of-ones.html
- Black Box Software Testing Foundations, Association for Software Testing, http://www.associationforsoftwaretesting.org
- Context Driven School of Testing, http://www.context-driven-testing.com
- Software Metrics: A Rigorous and Practical Approach by Norman E. Fenton and Shari Pfleeger
- Aristotle's Ethics, http://plato.stanford.edu/entries/aristotle-ethics/
- Social Research Methods, http://www.socialresearchmethods.net/