michael@developsense.com

Critical Thinking About Numbers and Measurement

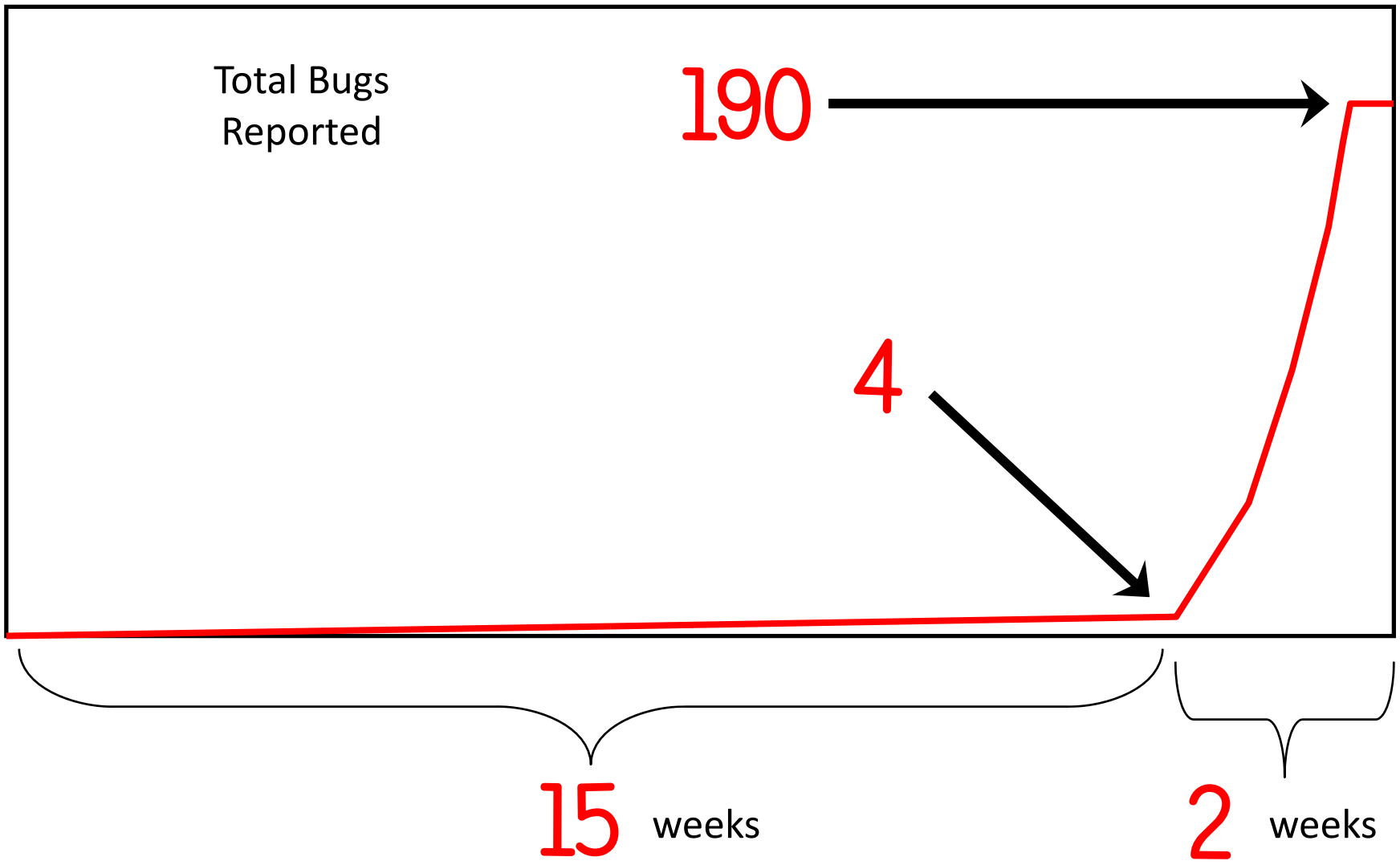
Michael Bolton
DevelopSense

Preface: I Don't Hate Numbers!

- I love numbers *so much* that I can't stand to see them abused as they are by people in our profession.
- This workshop is designed to take you deeper into measurement, spotting critical thinking errors that might cause you to miss observations and mislead your client—or yourself.
- The intention is not only to suggest that measurement has problems, but also to expand our notions of what good measurement might be.

> Imperfections in measurement are always a problem, but they're a devastating problem only when we don't recognize them.
>
> —Daniel Gilbert, Stumbling on Happiness

Example: A Test Project



Example: A Test Project

- The original tester was faking it.
- A tiger team took over for the last 2 weeks.
- The reporting system broke down in the last several days before the beta was shipped.
  (That's why the line goes flat–reports were still being prepared, but were not being formally logged in the tracking system.)

Lessons:

- Pay attention to the whole story.
- Don't assume that all known problems are being reported.
- *Think critically about the numbers.*

Exercise: Evaluating Claims

- Choose a claim from the next slide, and write it down.
  - Using a plausibility scale of 0-100 (where 0 is ridiculous and 100 is absolutely true), what is your assessment of the claim?
  - What is your thought process on encountering the claim?
  - What might change your evaluation of the claim?
  - Irrespective of your evaluation of the claim, what evidence would change your mind *to the opposite polarity*?
    - That is, if you disbelieved the claim, what could make you believe it? If you believed it, what could make you disbelieve it?

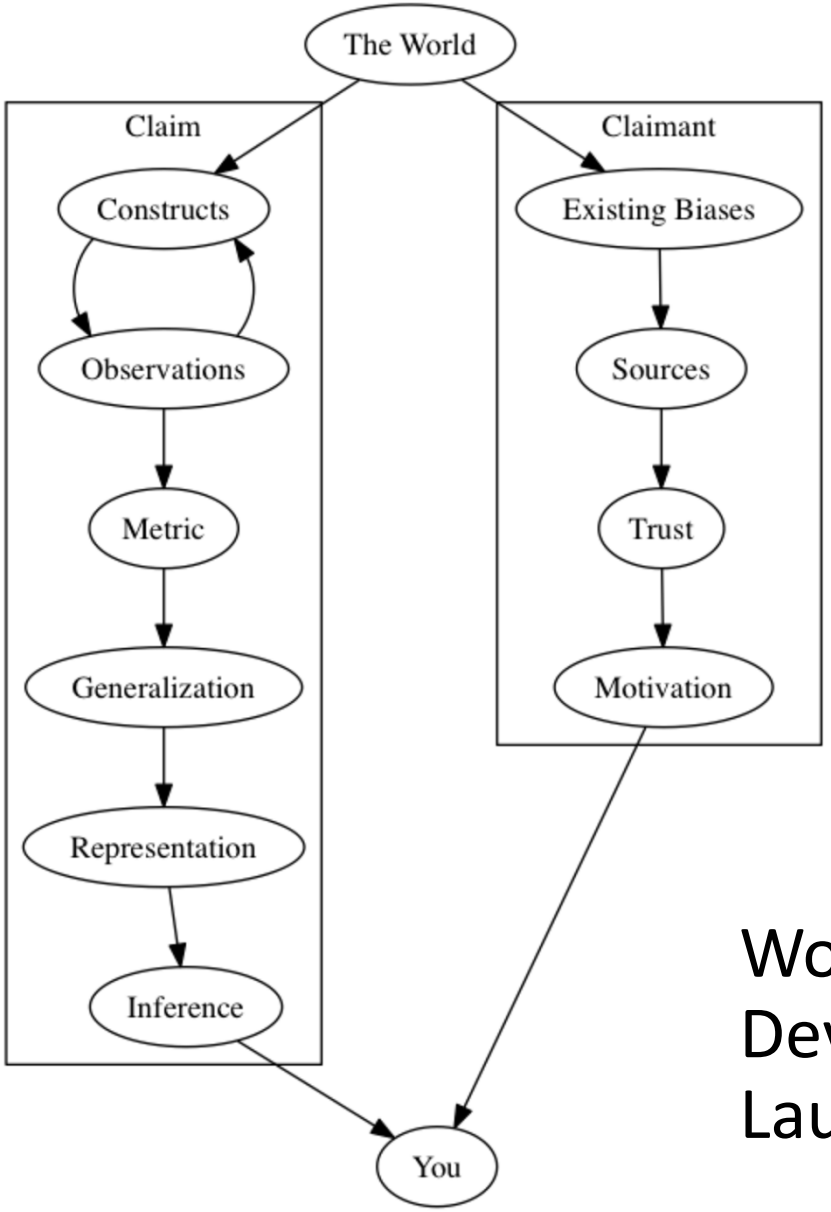
The Claims

- The average cost of correcting a bug during the coding phase is $977.
- Time and motion studies of actual defect repairs shows that fixing most bugs in code requires about four hours, regardless of whether the bug is found before or after release.
- Defect Removal Efficiency averages vary from 73% to 96% depending on organizations.
- The cost of defects rises exponentially the later they are detected.
- Software defects cost the US economy $60Bn annually.
- 25% of total defects are from bad fixes.
- The average time to find and fix a defect is 10 to 20 hours.

The Assignment

- Using a plausibility scale of 0-100 (where 0 is ridiculous and 100 is absolutely true), what is your assessment of the claim?
- What is your thought process on encountering the claim?
- What might change your evaluation of the claim? How would you test your belief?
- Irrespective of your evaluation of the claim, what evidence would change your mind *to the opposite polarity*?
  - That is, if you disbelieved the claim, what could make you believe it? If you believed it, what could make you disbelieve it?

A Model for Evaluating Claims



Work still in progress!
Developed in collaboration with Laurent Bossavit

One Event or Two?

Image Credit: www.theatlantic.com

Steven Pinker

The Stuff of Thought: Language as a Window into Human Nature

Is a Burrito a Sandwich?

Roy Sorenson
A Cabinet of Philosophical Curiosities

Image Credit: www.dreamstime.com

What is critical thinking?

The Nature of Critical Thinking

- "Critical thinking is **purposeful, self-regulatory judgment** which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based." *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction, Dr. Peter Facione*

(Critical thinking is, for the most part, about getting all the benefits of your "System 1" thinking reflexes while avoiding self-deception and other mistakes—including overdependence on System 2.)

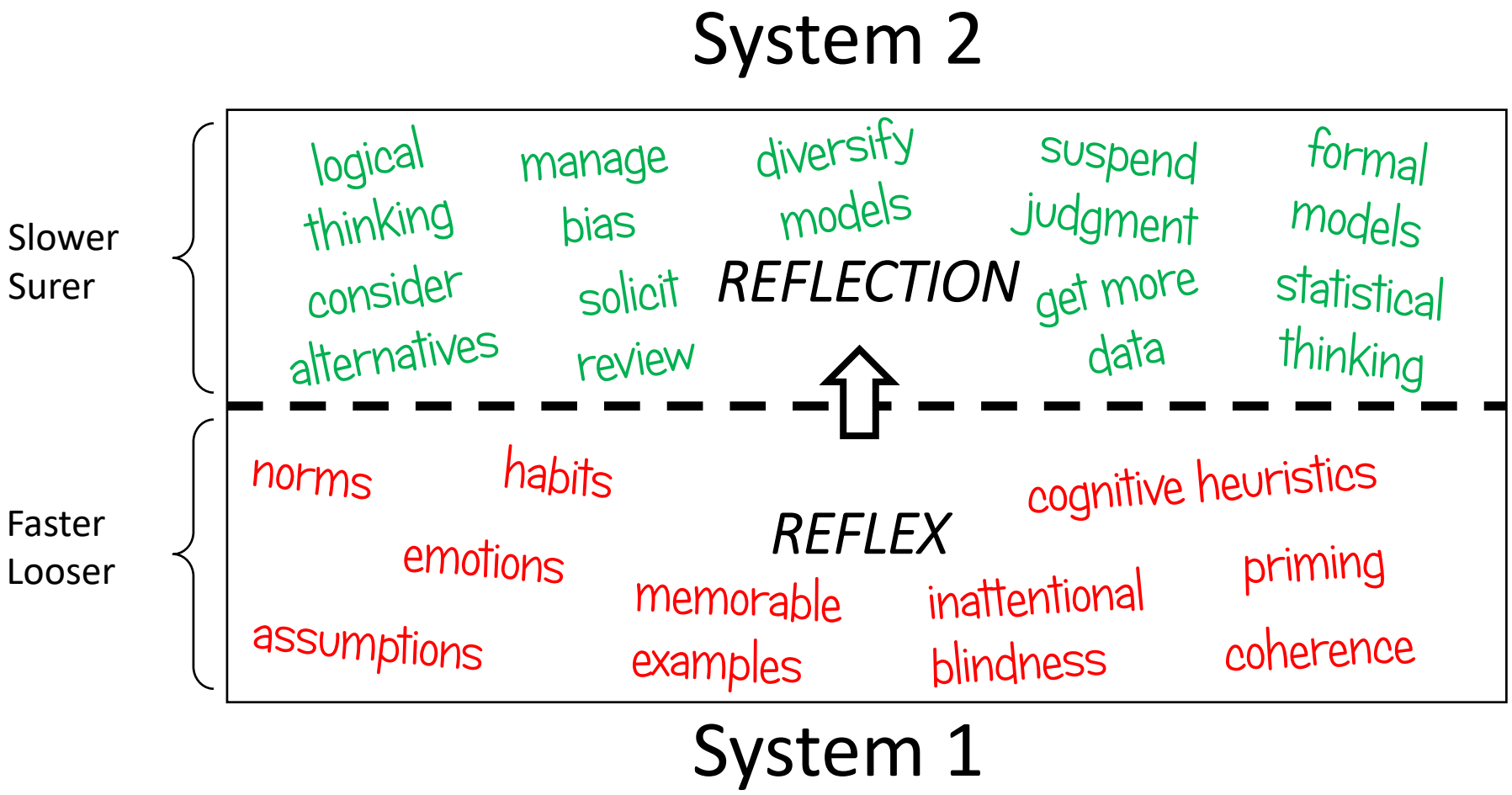
Bolton's Definition of Critical Thinking

Critical Thinking
is thinking about thinking
with the aim of not getting fooled.

- Michael Bolton

Testing is enactment of critical thinking about software to help people make better decisions.
Critical thinking must begin with our belief in the likelihood of errors in our thinking.

Reflex is IMPORTANT
But Critical Thinking is About Reflection



See Thinking Fast and Slow, by Daniel Kahneman

Workarounds for Our Bugs: Introducing Pauses

Giving System 2 time to wake up!

Huh?
- You may not understand. (errors in interpreting and modeling a situation, communication errors)

Really?
- What you understand may not be true. (missing information, observations not made, tests not run)

And?
- You may not know the whole story. (perhaps what you see is not all there is)

So?
- The truth may not matter, or may matter much more than you think. (poor understanding of risk)

Qualification Quips: Safety and Possibilities

- proportional principle: "to some degree"
- relative rule: "to some person, at some time"
- context concept: "in some context"
- uncertainty umbrella: "probably but not certainly"
- necessity nudge: "necessary but not sufficient"
- heuristic heuristic: "the solution is a heuristic“
- model modifier: “not the thing, but our model of the thing”
- particularity premise: “true for us in the here and now”
- debate delay: “we will decide on this later”
- design deferral: “we will solve this problem later”
- the *et cetera* escape: “there may be more to this”

Exercise
Apply critical thinking to this statement:

“Management wants numbers.”

What is measurement?

What Is Measurement?

"Measurement is
the empirical, objective assignment of numbers,
according to a rule derived from a model or theory,
to attributes of objects or events
with the intent of describing them."
—*Cem Kaner and Walter P. Bond*

Source: "Software Engineering Metrics: What Do They Measure and How Do We Know?" (Cem Kaner and Walter P. Bond)

http://www.kaner.com/pdfs/metrics2004.pdf

What happens when we walk through that definition?
What do we need to think critically about?
What could go wrong?

Exercise

Identify and describe
factors in measurement.

Some measurement factors…

- Object or event
- Attribute(s) of that object or event
  - measured / unmeasured
  - observed / unobserved
- Measuring instrument(s)
- Metric (how numbers get assigned)
- Rule
- Model or theory
- Description
- Intention (inquiry or control?)
- Observation
- Observer

…and some more measurement factors…

- Scale
- Precision
- Accuracy
- Validity
- Reliability
- Sample size
- Sample selection
- Errors

Going deeper…

- People
  - observers, consumers, subjects…
- System
  - relationship of attributes, objects, and events to others
- Construct
  - how to count to one
- Representation
  - how the measurement is displayed and described
- Generalization
  - how observations and conclusions might apply outside this context
- Inferences
  - what conclusions we could draw from this measurement

Problems and Sources of Error

- Validity
  - operationalization of constructs ("how to count to one")
  - relationship between what we're measuring and *what we think* we're measuring
  - degree to which we've accounted for alternative interpretations for our observations and conclusions
- Reliability
  - variations in attributes, instruments, observers
  - influenced by context: time, place, motivations…
- Biases and fallacies
  - too many to list here! See (e.g.) Wikipedia
- Side effects
  - distortion and dysfunction
  - people will often behave to optimize things that are being measured, at the expense of things that are not

Objectivity, Reliability, Validity

- *Objectivity* is the simultaneous realization of as much reliability and validity as possible.
- *Reliability* is the degree to which the finding is independent of accidental circumstances of the research
- *Validity* is the degree to which the finding is interpreted in a correct way.

Kirk, Jerome, and Miller, Mark, *Reliability and Validity in Qualitative Research*

Three Important Kinds of Reliability

- Quixotic reliability
  - the measurement yields consistent results in different circumstances
  - could be the result of a broken instrument, or socially acceptable answers
- Diachronic reliability
  - the measurement yields consistent results when taken multiple times
  - only reliable for things that don't change in a changing world; "may deny history"
- Synchronic reliability
  - similarity of observations within the same time period
  - reveals potentially significant questions when it fails

Kirk, Jerome, and Miller, Mark, *Reliability and Validity in Qualitative Research*

Construct Validity & External Validity

- *Construct* validity is (informally) the degree to which your attributes and measurements can justified within an experiment or observation
  - How do you demarcate the difference between *one* of something and *not-one* of something?
  - How do you know that you're measuring what you think you're measuring?
- *External* validity is the degree to which your experiment or observation can be generalized to the world outside
  - How do you know that your experiment or observation will be relevant at other times or in other places?

"In the case of qualitative observations, the issue of validity is not a matter of methodological hair-splitting about the fifth decimal point, but a question of whether the researcher sees what he or she thinks he or she sees."

Kirk, Jerome, and Miller, Mark, *Reliability and Validity in Qualitative Research*

Exercise

Analyse Kaner and Bond's definition.

How can we sharpen our analysis of a given measurement?

Kaner & Bond on Measurement Validity

1. What is the purpose of your measurement?
2. What is the scope of the measurement?
3. What is the attribute you are trying to measure?
4. What are the scale and variability of this attribute?
5. What is the instrument you're using
6. What are the scale and variability of the instrument ?
7. What function (metric) do you use to assign a value to the attribute?
8. What's the natural scale of the metric?
9. What is the relationship of the attribute to the metric's output?
10. What are the natural, foreseeable side effects of using this measure?

The essence of good measurement is a model that incorporates answers to questions like these.

If you don't have solid answers, you aren't doing measurement; you are just playing with numbers.

Exercise

Using Kaner and Bond's questions,
analyze this statement:

"Defect Detection Efficiency is a good way to evaluate the performance of the testing group."

Critical Thinking About Measurement

from an actual client

DER – Defect Escape Rate

The Defect Escape Rate measures the number of undiscovered defects that escaped detection in the product development cycle and were released to customers. An escape is a defect found while using a released product. DER is defined as:

$$DER = (Defect\ Escapes\ /Total\ Defects) * 100$$

DER is a lagging indicator of product quality. The number of escapes is always zero until after product is released. It is reported as a percentage and a low number is desired. Each business unit has a target DER percentage and an Escape Analysis should be performed on each defect to improve test coverage. It is desirable for the DER for a *product line* to decline over time. See appendix for calculation details.

Look! Measurement!

What could possibly go wrong?

An Alternative View of Measurement

Measurement is the art and science
of making reliable and significant observations.
—Jerry Weinberg, *Quality Software Management Vol. 2*

- Since the time of Aristotle (at least), we've known about two kinds of measurement that inform decisions
  - "Two pounds of meat"
  - "Too much", "too little", "just right".

We waste time and effort when we try to obtain
six-decimal-place answers to whole-number questions.

- http://www.developsense.com/articles/2009-05-IssuesAboutMetricsAboutBugs.pdf
- http://www.developsense.com/articles/2009-07-ThreeKindsOfMeasurement.pdf
- http://www.developsense.com/articles/2007-11-WhatCounts.pdf

How Do We Measure?

- Third-order measurement
  - highly instrumented, used to discover natural laws
  - "What *will* happen? What *always* happens?"

- Second-order measurement
  - often instrumented, used to refine first-order observation
  - used to tune existing systems
  - "What's *really* going on here? What's happening right now?"

How Else Do We Measure?

- First-order measurement
  - minimal fuss, direct observation, minimal instrumentation
  - used to inform a control action OR to prompt search for more refined information
  - “What’s going on? What should we do? Where should we look?”

Weinberg suggests that, in software development, we’re obsessed with trying to make third- and second-order measurements when first-order measurements might be all we need—and tend to be much cheaper and easier.

Why Prefer First-Order Measures?

- When you're driving, are you mostly concerned about…
  - your velocity, acceleration, vehicle mass, drag co-efficient, frictional force? (third-order)
  - your engine temperature, RPMs, and current rate of gas consumption? (second-order)
  - looking out the window to avoid hitting things (first-order)?

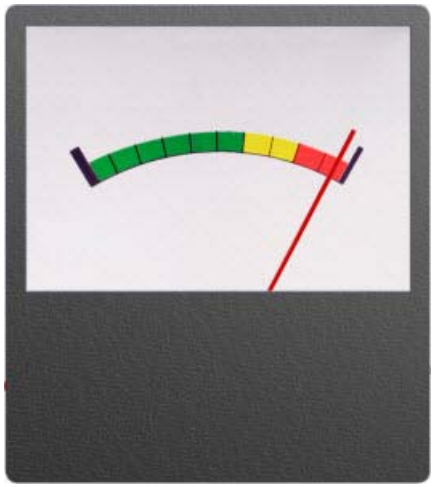
I've observed *many* projects that have crashed because managers were overly focused on the dashboard instead of the traffic and obstacles around them, and the road ahead.

What kind of driver do you trust?

Control vs. Inquiry Measurement

- A **control measurement** is a measurement that **drives decisions**.
  - *Any measurement you use to control a self-aware system will be used by that system to control YOU.*

- An **inquiry measurement** is any measurement that **helps you ask the right questions at the right time**.
  - Inquiry measurements are also vulnerable to gaming, but the stakes are far lower, so there's less incentive for manipulation.



Text here is taken from the work of my colleague, James Bach.
http://www.satisfice.com

Control vs. Inquiry

- Remove *control metrics* that are linked to pay, bonuses, performance evaluation, etc.
  - control metrics trigger some action, usually automatically
  - a metric that is used to control something will eventually be used to control you
- Foster *inquiry metrics*
  - inquiry metrics prompt us to ask questions
- Relax measurement when the metric stops changing
  - if you're not obtaining new information, try measuring something else of interest for a while

"But they ask us for numbers!"

What are they *really* asking for?

- “We want to know if we’re improving.”
- “We want to know if we’re happy.”
- “We want to know if we should be *un*happy.”

- Are they asking for *specific* numbers?
  - Perform the Kaner/Bond checklist
  - Offer a list of threats to validity
  - Provide a number, and include commentary

"But they ask us for numbers!"

Are they asking for *numbers*?

- Offer an *observation*
  - e.g. "cold enough to freeze the water in the bird bath"
- Offer a *description*
  - e.g. a product status report; a coverage outline
- Offer a *list*
  - e.g. a list of problems in the product; a list of project problems
- Offer a *table*
  - e.g. time spent on classes of activities
- Offer a *visual model*
  - e.g. diagrams of effects, Wiggle charts, mind maps…
- Offer a comparison or ranking

"But they ask us for numbers!"

- Prefer first-order measurement
- Prefer measurement for *inquiry* to measurement for *control*
- Ask “compared to what?”

What *could* we measure?

Degrees of Coverage

Level 0	**We don't really know anything about this area.** We're aware that this area exists, but it's a black box to us, so far.
Level 1	**We're just getting to know this area.** We've done basic reconnaissance; surveyed it; we've done smoke and sanity testing. We may have some artifacts that represent our models, which will helps us to talk about them and go deeper.
Level 2	**We've learned a good deal about this area.** We've looked at the core and the critical aspects of it. We've done some significant tests focused on the most important quality criteria, and we're collecting and diversifying our ideas on how to cover it deeply.
Level 3	**We have a comprehensive understanding of this area.** We've looked deeply into it from a number of perspectives, and applied a lot of different test techniques. We've done harsh, complex, and challenging tests on a wide variety of quality criteria. If there were a problem or unrecognized feature in this area that we didn't know about, it would be a big surprise.

Time Spent on Testing Work

Testing (T)	Active test design; experimentation, interaction, learning about the product; increasing test coverage.
Bug (B)	Study and investigation of bugs; finding repro steps; looking for similar bugs inside a session. B–time interrupts T-time.
Setup (S)	Work within a session to prepare for testing, to support it, **or to follow up on it**. Setting up products, tools, environments; studying; analyzing non-bug behaviour… S-time interrupts T-time.
Opportunity	Work within a session that is NOT directed towards fulfilling the charter, but towards the general mission of testing. Chasing after a risk, helping other testers, testing while waiting for something else to happen…
Non-session	Meetings, lunches, breaks, chat, work-related or personal business done outside of a testing session.

You can’t measure quality…
but you can discuss it.

—James Bach